**模型的存储问题和模型进行预测的速度问题**

卷积神经网络（简称 CNN）在图像分类、图像分割、目标检测等领域获得广泛应用。随着性能要求越来越高，AlexNet 已经无法满足大家的需求，于是乎各路大牛纷纷提出性能更优越的 CNN 网络，如 VGG、GoogLeNet、ResNet、DenseNet 等。由于神经网络的性质，为了获得更好的性能，网络层数不断增加，从 7 层 AlexNet 到 16 层 VGG，再从 16 层 VGG 到 GoogLeNet 的 22 层，再到 152 层 ResNet，更有上千层的 ResNet 和 DenseNet。虽然网络性能得到了提高，但随之而来的就是效率问题。

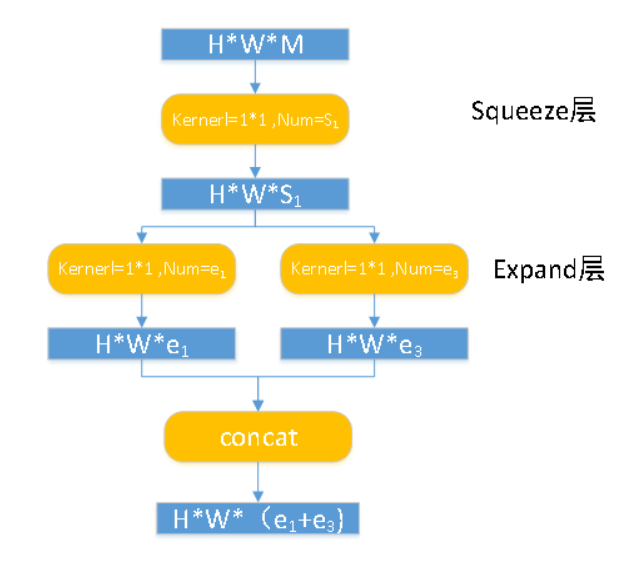
**四个轻量化模型**

SqueezeNet、MobileNet、ShuffleNet、Xception

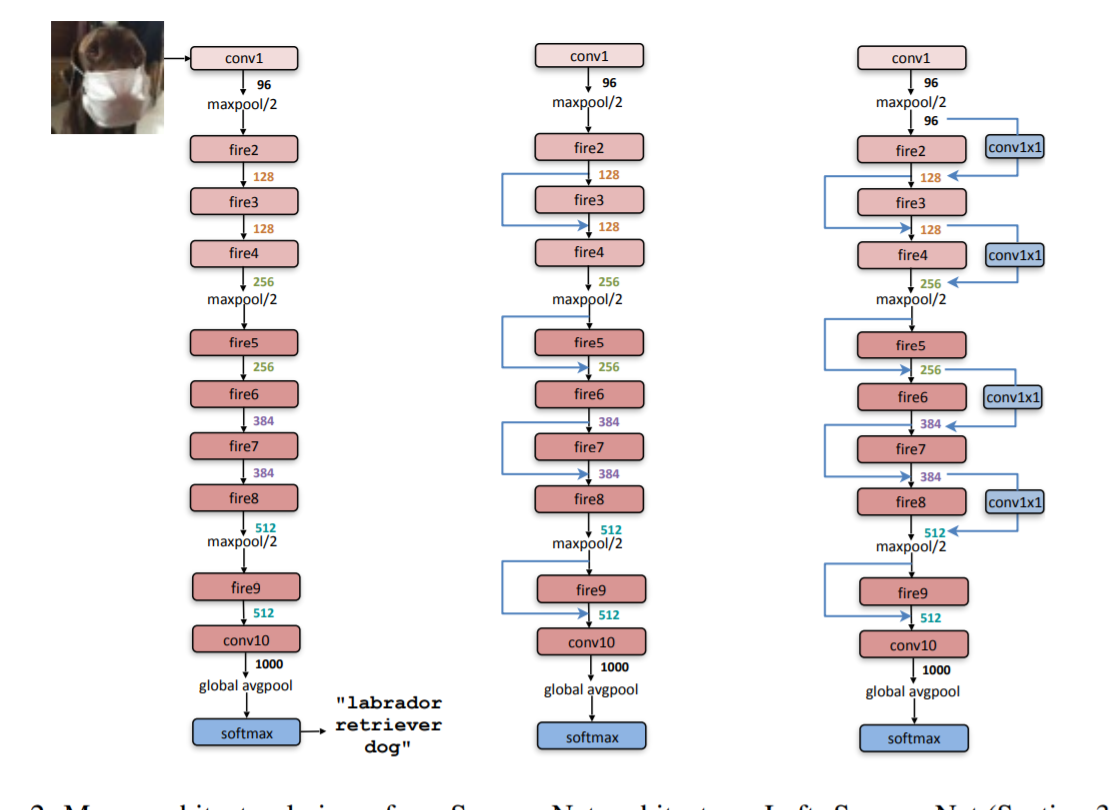
**这四种轻量化模型仅是在卷积方式上做了改变**

**SqueezeNet（SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND <0.5MB MODEL SIZE）（1DeepScale∗ & UC Berkeley 2Stanford University）**

采用不同于传统的卷积方式，提出 fire module；fire module 包含两部分：squeeze 层+expand 层



**在原文中使用，e1=e3=4s1**

Fire module 与 GoogLeNet 思想类似，采用 1\*1 卷积对 feature map 的维数进行「压缩」，从而达到减少权值参数的目的；

**MobileNet（MobileNets: Efficient Convolutional Neural Networks for Mobile Vision**

**Applications）**