细粒度视觉分类（FGCV，Fine-Grained Visual Categorization）即识别细分类别的任务，一般它需要同时使用全局图像信息与局部特征信息精确识别图像子类别。细粒度分类是计算机视觉社区最为有趣且有用的开放问题之一，目前还有很多难题期待解决。

由于子类别间细微的类间差异以及较大的类内差异，较之普通的图像分类任务，细粒度图像分类难度更大。 图1所示为细粒度图像分类数据集CUB-200[1]中的两个物种，加州鸥和北极鸥，从竖直方向的图片对比可以看出，两个不同物种长相非常相似，而从对比水平方向可知，同一物种由于姿态，背景以及拍摄角度的不同，存在较大的类内差异。因此，要想顺利的对两个极为相似的物种进行细粒度分类，最重要的是在图像中找到能够区分这两个物种的区分性的区域块(discriminative part)，并能够对这些有区分性的区域块的特征进行较好的表示。

从竖直方向的图片对比可以看出，两个不同物种长相非常相似，而从对比水平方向可知，同一物种由于姿态，背景以及拍摄角度的不同，存在较大的类内差异。 因此，要想顺利的对两个极为相似的物种进行细粒度分类，最重要的是在图像中找到能够区分这两个物种的区分性的

区域块(discriminative part)，并能够对这些有区分性的区域块的特征进行较好的表示。

由于深度卷积网络能够学习到非常鲁棒的图像特征表示，对图像进行细粒度分类的方法，大多都是以深度卷积网络为基础的，这些方法大致可以分为以下四个方向：

1. 基于常规图像分类网络的微调方法
2. 基于细粒度特征学习的方法（fine-grained feature learning)
3. 基于目标块的检测（part detection)和对齐(alignment)的方法
4. 基于视觉注意机制（visual attention)的方法

一、基于常规图像分类网络的方法

这一类方法大多直接采用常见的深度卷积网络来直接进行图像细粒度分类，比如AlexNet[3]、VGG[4]、GoogleNet[5]、ResNet[6]以及DenseNet[7]和SENet[8] 等。

由于这些分类网络具有较强的特征表示能力，因此在常规图像分类中能取得较好的效果。然而在细粒度分类中，不同物种之间的差异其实十分细微，因此，直接将常规的图像分类网络用于对细粒度图像的分类，效果并不理想。受迁移学习理论启发，一种方法是将大规模数据上训练好的网络迁移到细粒度分类识别任务中来。常用的解决方法是采用在ImageNet上预训练过的网络权值作为初始权值，然后再通过在细粒度分类数据集上对网络的权值进行微调（finetune），得到最终的分类网络。

在[9]中，Zhang等人进一步将度量损失函数引入到精细分类网络的微调中来。具体而言，每次输入三个样本（Postive，Reference以及Negative）到三个共享权值的网络中，然后利用三个网络的特征输出用来计算损失函数，除了传统的softmax 损失函数，三个特征输出还构成了广义的triplet 损失。最后两个损失函数联合用来微调网络:

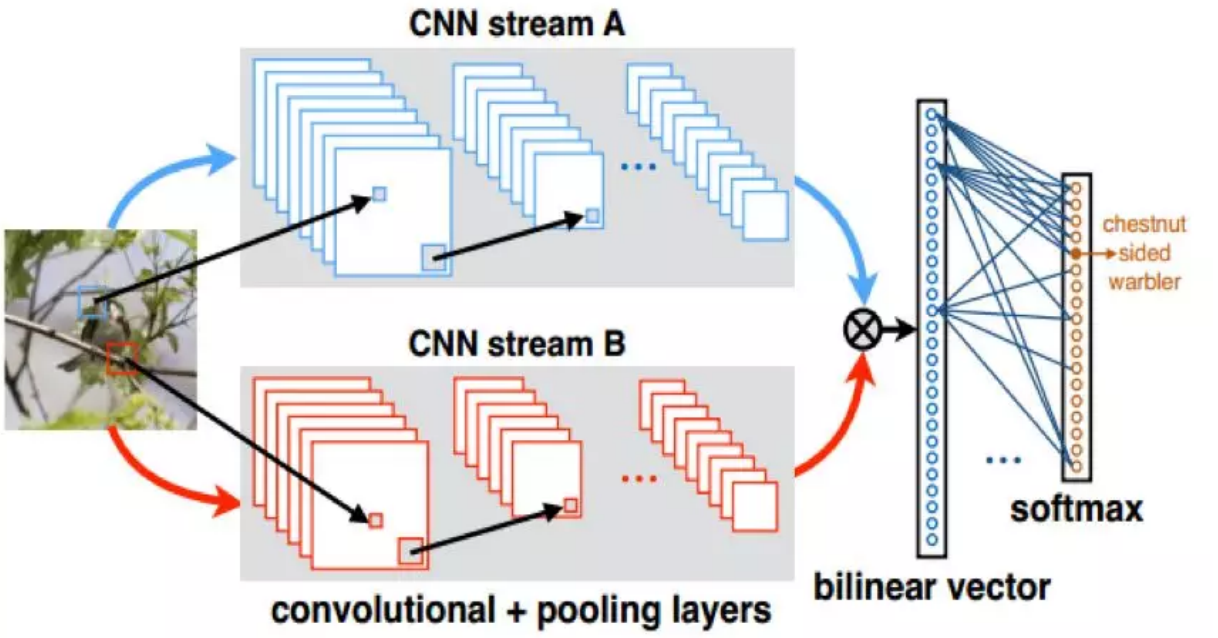
$$E = \lambda_s E_s(r) + (1-\lambda) E_t(r,p,n)$$

其中，$E_S(r)$是softmax获取的分类误差，$E_t(r,p,n)$是通过图2中三个共享参数的子网络$f_r(s)$，$f_p(s)$和$f_n(s)$获取到的triplet误差，两种误差实现对网络不同层次的约束。$E_S(r)$通过图像的类别信息，约束网络参数的优化方向是在图像真实类别上获取最大的响应，这其中

并没有关注不同类别之间的度量关系。而$L_t(r,p,n)$则通过计算类内距离与类间距离，增大网络对不同类别的相似样本的识别能力。

二、基于细粒度特征学习的方法

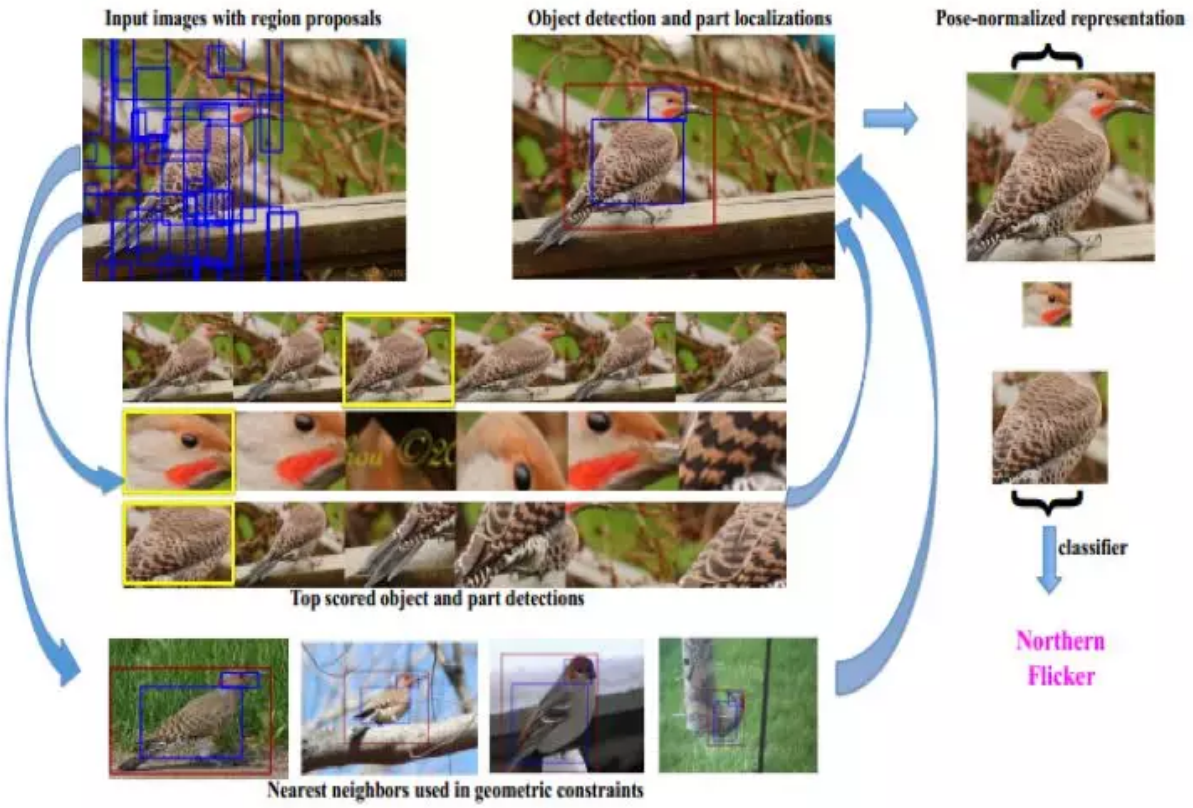
Lin等人在2015年发表于ICCV的论文[9]中提出双线性卷积神经网络模型（Bilinear CNN，网络结构如图3所示）实现对深度卷积特征更好的表示。该方法使用VGG-D和VGG-M两个网络作为基准网络，在不使用Bounding Box （边框）标注信息的情况下，在CUB200-2011数据集上到达了84.1%的分类精度；而使用Bounding Box时，其分类精度高达85.1%。



总体来说，双线性CNN模型能够基于简洁的网络模型，实现对细粒度图像的有效识别。一方面，CNN网络能实现对细粒度图像进行高层语义特征获取，通过迭代训练网络模型中的卷积参数，过滤图像中不相关的背景信息。更重要的是另一方面，网络A 和网络B在图像识别任务中扮演着互补的角色，即网络A能够对图像中的物体进行定位,, 而网络B 则是完成对网络A 定位到的物体位置进行特征提取。通过这种方式，两个网络能够配合完成对输入细粒度图像的类检测和目标特征去的过程，较好地完成细粒度图像识别任务。关于双线性网络更加的介绍可以参考SIGAI的另外一篇文章：双线性汇合(bilinear pooling)在细粒度图像分析及其他领域的进展综述。

三、基于目标块检测的方法

基于目标块（object part）检测的方法思路是：先在图像中检测出目标所在的位置，然后再检测出目标中有区分性区域的位置，然后将目标图像（即前景）以及具有区分性的目标区域块同时送入深度卷积网络进行分类。但是，基于目标块检测的方法，往往在训练过程中需要用到目标的Bounding box标注信息，甚至是目标图像中的关键特征点信息，而在实际应用中，要想获取到这些标注信息是非常困难的。比较有代表性的是2014年ECCV中提出来的Part-RCNN方法[11]：



四、基于注意力机制

由于基于视觉注意模型(Vision Attention Model)的方法可以在不需要额外标注信息（比如目标位置标注框和重要部件的位置标注信息）的情况下，定位出图像中有区分性的区域，近年来被广泛应用于图像的细粒度分类领域。代表性的工作是17年CVPR中提出的循环注意卷积神经网络(Recurrent Attention Convolutional Neural Network, RA-CNN)[14]。该模型模仿faster-RCNN[15]中的RPN (Region Proposal Network) 网络，提出使用APN (Attention Proposal Network) 网络来定位出图像中的区分性区域，并通过在训练过程中使用排序损失函数 (Rank Loss)，来保证每次利用注意模型定位的区域都更加有效。

一般细粒度识别可以分为两种，即基于强监督信息的方法和仅使用弱监督信息的方法。。前者需要使用对象的边界框和局部标注信息，后者仅使用类别标签，例如 2014 年提出的 Part-based R-CNN 利用自底向上的候选区域（region proposals）计算深度卷积特征而实现细粒度识别。这种方法会学习建模局部外观，并加强局部信息之间的几何约束。而 iMaterialist 2018 仅使用类别标签，因此是一种弱监督信息的细粒度识别。

码隆科技首席科学家黄伟林博士总结，在多年从事商品识别的研究和实践过程中，面临的三个主要难点。首先，细粒度商品识别，特别是对 SKU 级别的识

别是至关重要的。如下图所示，不同种类的益达口香糖，在零售过程中通常价格会不太一样，因此需要作精确区分。其次，除了细粒度分析，SKU 级别的商品识别通常需要识别大量的商品种类，比如超过 10 万类，而常见的 ImageNet 物体识别通常只有 1,000 类。这是商品识别的另一个挑战，而常用的单层 softmax 分类模型很难解决。
这就需要引进多层级联的细粒度分类算法，从而加大细粒度识别的难度。最后，由于商品类别多，就要去更多的海量训练数据和人工标注，比如 10 亿级别的。对于如此数量的人工标注和数据清洗，是很难完成的。因此，如何有效地利用海量网络爬去的商品图片，在没有或者只有少量人工标注和清洗的情况下，训练一个高性能的商品识别模型，成为一个关键的技术。码隆科技最近提出的弱监督学习算法- CurriculumNet，就是专门为训练海量无工人共标注的海量网络图片而设计的。

解决细粒度图像分析的一个关键是找到细粒度物体的Keypoints，利用这些关键部位的不同，进行针对性的细粒度分析，如检索、识别等。目前，细粒度图像分析领域的经典基准数据集包括：

- 鸟类数据集CUB200-2011，11788张图像，200个细粒度分类
- 狗类数据集Stanford Dogs，20580张图像，120个细粒度分类
- 花类数据集Oxford Flowers，8189张图像，102个细粒度分类
- 飞机数据集Aircrafts，10200张图像，100个细粒度分类
- 汽车数据集Stanford Cars，16185张图像，196个细粒度分类

Mask-CNN细粒度图像识别基本思想

细粒度图像识别也可以被称为细粒度图像分类。在进行细粒度分类时，研究人员借用了做图像检索时的思想，即“定位主要物体，去掉无效描述子（describetor）”。后续实验可证实，基于深度描述子筛选的思想不仅有利于细粒度图像检索，对细粒度图像识别同样大有裨益。

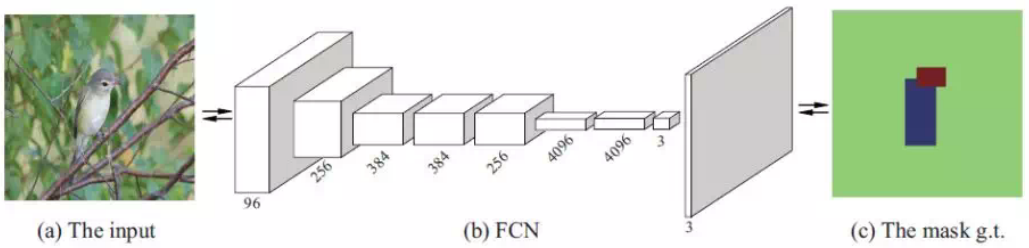
具体而言，研究人员通过对分析对象添加keypoints，进而生成keyparts，从而生成有关ground truth的边界框。

Part-based Mask-CNN

(a) Part annotations　　(b) Part rectangles

然后，研究人员用生成的边界框来训练一个分割网络，学习针对物体的mask。

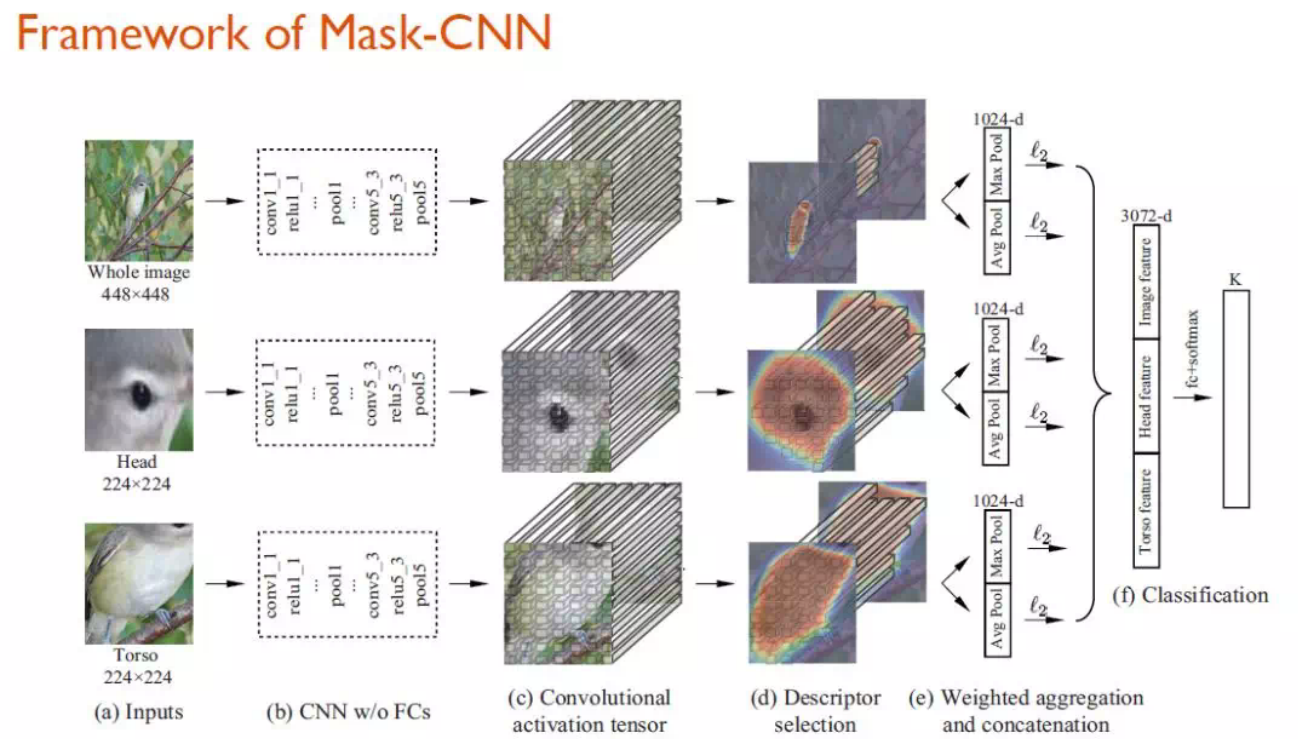
Learning part masks by FCN



(a) The input　　(b) FCN　　(c) The mask g.t.

Figure 3: Demonstration of the mask learning procedure by FCN. (Best viewed in color.)

进一步，研究人员根据习得的mask，生成了对应物体整体、头部、身体的三个网络。他们在这些网络的最后一层卷积tensor上进行对应的深度描述子筛选与融合。最后，通过将融合的结果进行级联，研究人员便可以让模型做最后的细粒度识别任务。



细粒度图像分析发展展望

少量样本细粒度图像识别

目前所有细粒度图像识别任务均需借助大量、甚至海量的标注数据。对于细粒度图像而言，其图像收集和标注成本巨大。如此便限制了细粒度研究相关的发展及其在现实场景下的应用。反观人类，我们则具备在极少监督信息的条件下学习新概念的能力，例如，对于一个普通成年人可仅借助几张图像便学会识别鸟类的一个新物种。为了使细粒度级别图像识别模型也能像人类一样拥有少量训练样本下的学习能力，研究人员提出并研究了细粒度级别图像识别的少量样本学习任务。