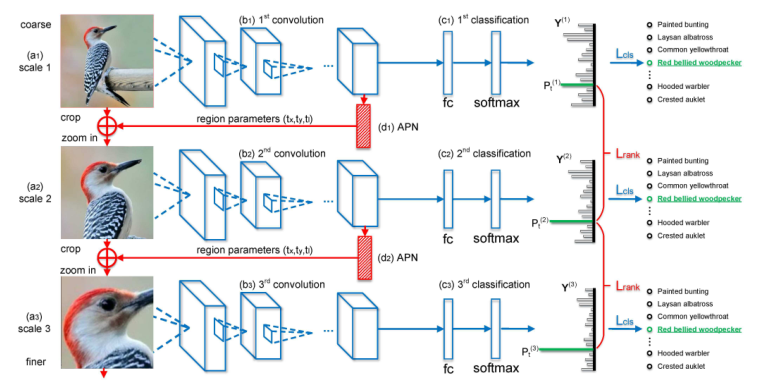
**[Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition] CVPR2017**

**主要思想：**RA-CNN在**三个尺度**上都有一个**分类子网络**和一个**APN**（attention proposal sub-network）。从完整图像开始，输出的区域预测作为下一个尺度的参考，由粗到细迭代生成区域注意力。训练损失有尺度内分类损失和尺度间排序损失，达到区域注意力和细粒度特征互相促进。不需要边界标注，端到端训练。

**贡献/缺点：**原有的细粒度物体种类识别（如鸟的种类或汽车型号）是在判别区域定位（discriminative region localization）和细粒度特征学习（fine-grained feature learning）两个层面单独解决的，本文提出循环注意力卷积神经网络（RA-CNN）递归地多尺度地学习判别区域定位和区域特征表示。

**论文细节：**



**网络结构。**网络先后在a1,a2,a3三个尺度上学习，a2,a3尺度上的输入分别是a1,a2输出的区域。每个尺度上采用的网络结构是相同的，都是multi-task且有**两个输出**，以a1为例，输入原图由CNN提取特征，最后的特征图输入全连接层和softmax层得到第一个输出即在细粒度种类上的概率分布(probability distribution p over fine-grained categories)。同时，特征图作为APN(只有两个全连接层，训练耗费资源非常小)的输入，得到第二个输出[tx,ty,tl]，即一组attention region 的坐标预测，tx,ty是正方形区域的中心坐标，tl是边长值的一半。基于这组坐标将原图crop并zoom in进入送入更小尺度提取更细粒度的特征。流程相似，网络结构相同，最后一个尺度只输出细粒度种类概率分布。（可堆叠更多的尺度）

**Attention mask.**为了保证APN在训练时可以被优化，本文将一个二维的阶跃函数作为attention mask来cropping。采用这个attention mask来crop而不是直接基于根据坐标和边长直接crop的原因是，端到端的优化训练需要引入连续函数。这个mask可以在前向传播时选择最重要的区域并在反向传播是可被优化（连续函数的性质）。Crop后的图像通过双线性插值缩放至网络所需尺寸。Mask具体构造方式见原文和链接<https://blog.csdn.net/qq_36302589/article/details/88664461>**。**

**Multi-task损失函数。** 训练损失是三个尺度的尺度内分类损失和尺度间排序损失的和。细节看原文。

