1. 什么是Attention机制？

其实我没有找到attention的具体定义，但在计算机视觉的相关应用中大概可以分为两种：

1）**学习权重分布：输入数据或特征图上的不同部分对应的专注度不同**，对此Jason Zhao在知乎回答中概括得很好，大体如下：

- 这个加权可以是保留所有分量均做加权（即soft attention）；也可以是在分布中以某种采样策略选取部分分量（即hard attention），此时常用RL来做。

- 这个加权可以作用在原图上，也就是《Recurrent Model of Visual Attention》（RAM）和《Multiple Object Recognition with Visual Attention》（DRAM）；也可以作用在特征图上，如后续的好多文章（例如image caption中的《 Show, Attend and Tell: Neural Image Caption Generation with Visual Attention》）。

- 这个加权可以作用在空间尺度上，给不同空间区域加权；也可以作用在channel尺度上，给不同通道特征加权；甚至特征图上每个元素加权。

- 这个加权还可以作用在不同时刻历史特征上，如Machine Translation。

2） **任务聚焦：通过将任务分解，设计不同的网络结构（或分支）专注于不同的子任务，重新分配网络的学习能力，从而降低原始任务的难度，使网络更加容易训练。**

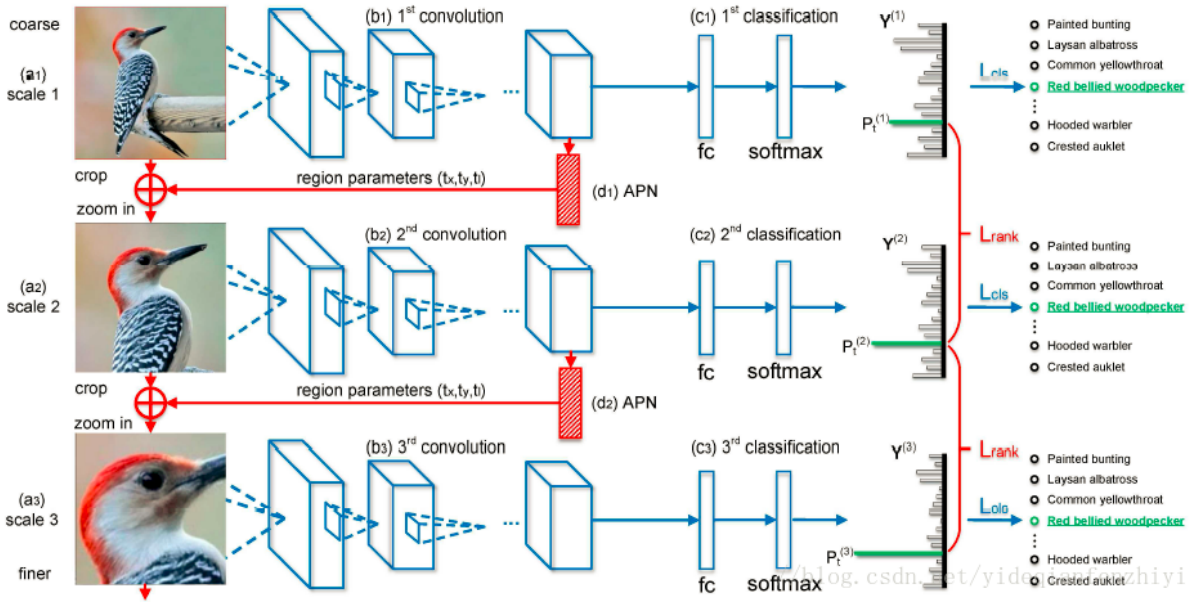
2. Attention机制应用在了哪些地方？

针对于1部分中的attention的两大方式，这里主要关注其在视觉的相关应用中。

2.1 方式一：学习权重分布

&1.（精细分类）Jianlong Fu, Heliang Zheng, Tao Mei (Microsoft Research), Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-grained Image Recognition, CVPR2017 （这个文章很有意思）

在关注的每一个目标尺度上，都采用一个分类的网络和一个产生attention proposal 的网络(APN)。本文最有趣的就是这个APN。这个APN由两个全连接层构成，输出3个参数表示方框的位置，接下来的尺度的分类网络只在这个新产生的方框图像中提特征进行分类。怎么训练呢？**本文定义了一个叫做rank Loss，用这个loss来训练APN，并强迫finer的尺度得到的分类结果要比上一个尺度的好，从而使APN更提取出更有利于精细分类的目标局部出来。**通过交替迭代训练，APN将越来越聚焦目标上的细微的有区分性的部分。当然这里有一个问题，那就是精细尺度只能聚焦到最显著的部位（如鸟头），但其他部分（如羽毛、鸟爪）就关注不到了。



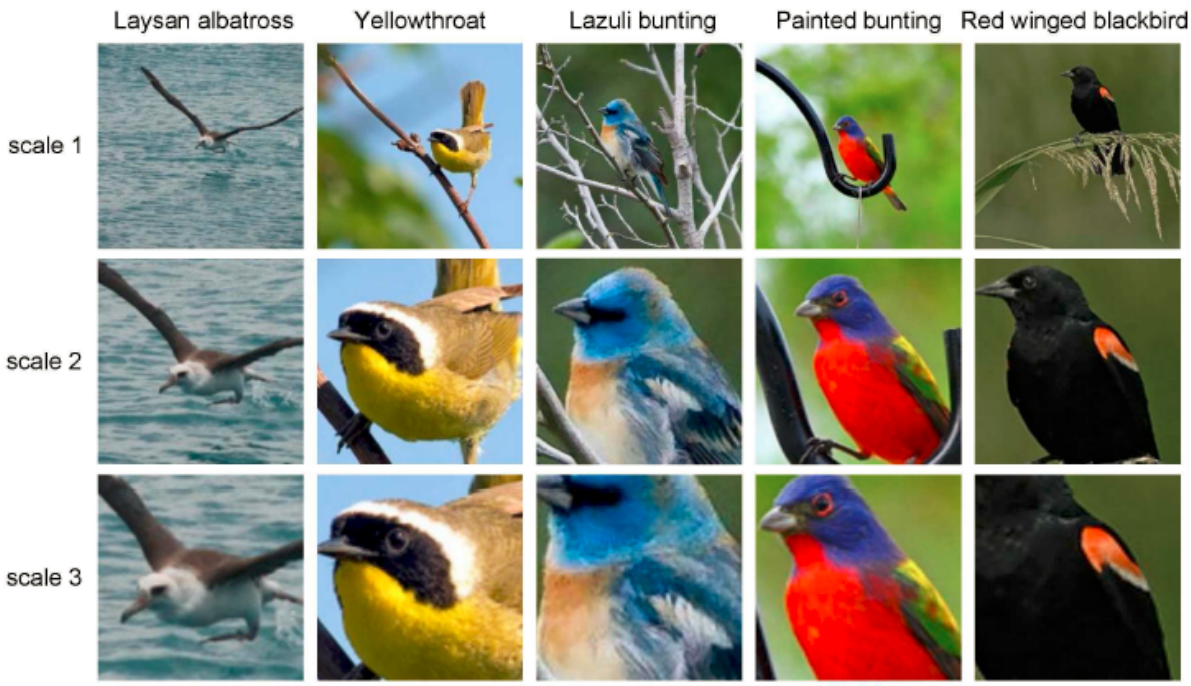
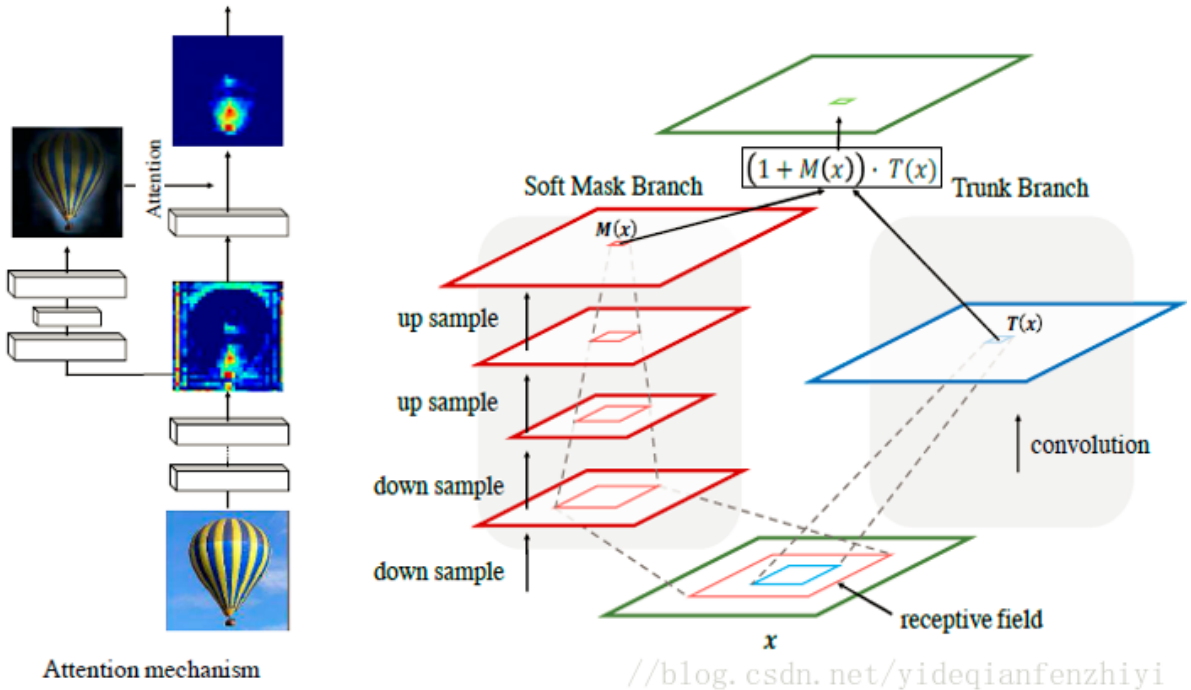


Figure 4. Five bird examples of the learned region attention at different scales. We can observe clear and significant visual cues for classification after gradually zooming in the attended regions.

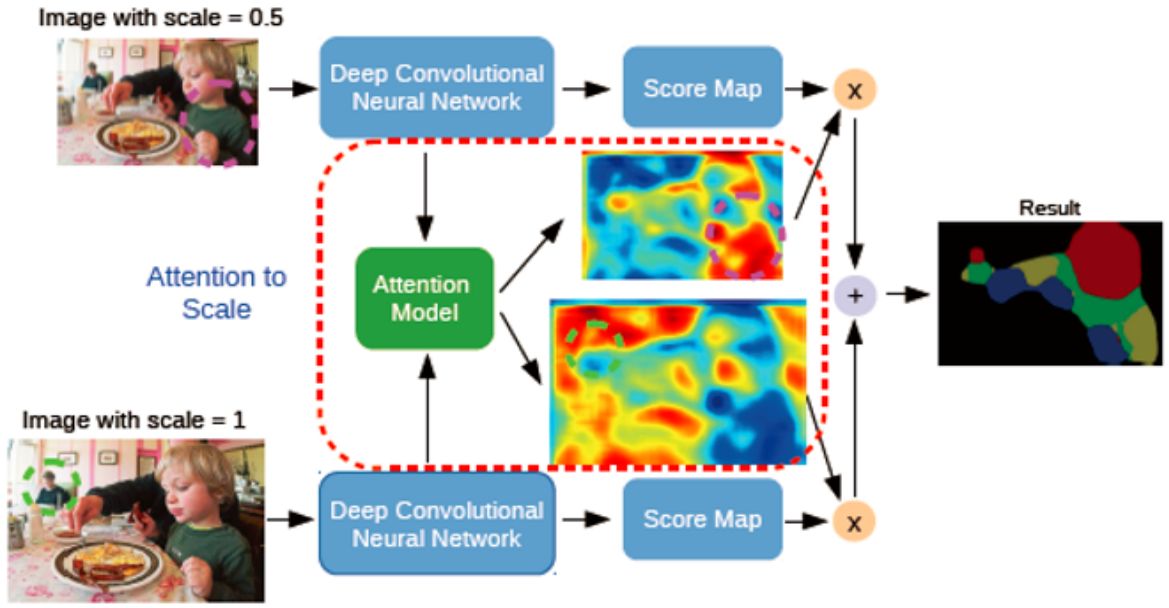
&2. （图像分类）Fei Wang, etc. (SenseTime Group Limited). Residual Attention Network for Image Classification，CVPR2017

本文是在分类网络中，增加了Attention module。这个模块是由两支组成，一支是传统的卷积操作，另一支是两个下采样加两个上采样的操作，目的是获取更大的感受野，充当attention map。因为是分类问题，所以高层信息更加重要，这里通过attention map提高底层特征的感受野，突出对分类更有利的特征。相当于变相地增大的网络的深度。



&3.（图像分割）Liang-Chieh Chen，etc.（UCLA）Attention to Scale: Scale-aware Semantic Image Segmentation, CVPPR2016（权重可视化效果有点意思）

通过对输入图片的尺度进行放缩，构造多尺度。传统的方法是使用average-pooling或max-pooling对不同尺度的特征进行融合，而**本文通过构造Attention model（由两个卷积层构成）从而自动地去学不同尺度的权重，进行融合（效果提升1到2个点吧，不同的数据集不一样）**。从论文中的权重可视化的结果，能发现大尺寸输入上，对应网络关注于small-scale objects，而在稍微小一点的尺寸输入上，网络就关注于middle-scale，小尺寸输入则关注background contextual information。可视化效果感觉非常有意思。



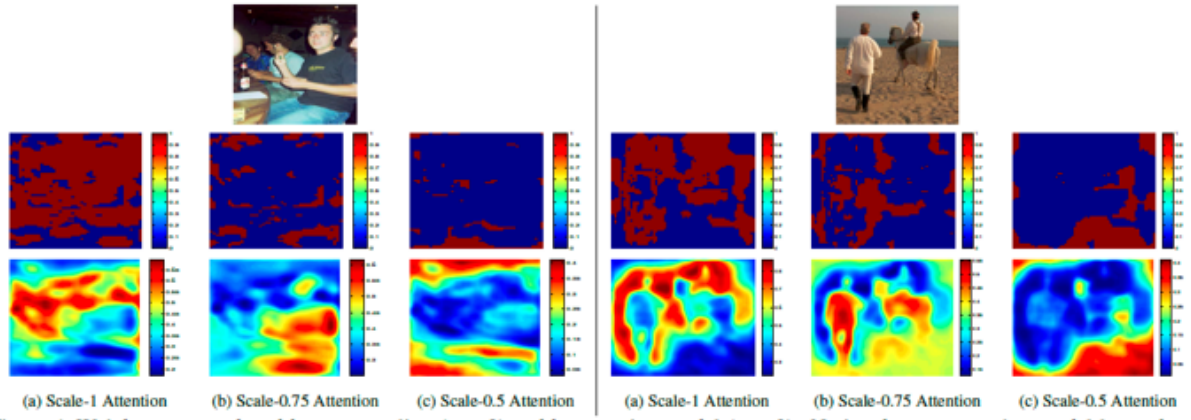
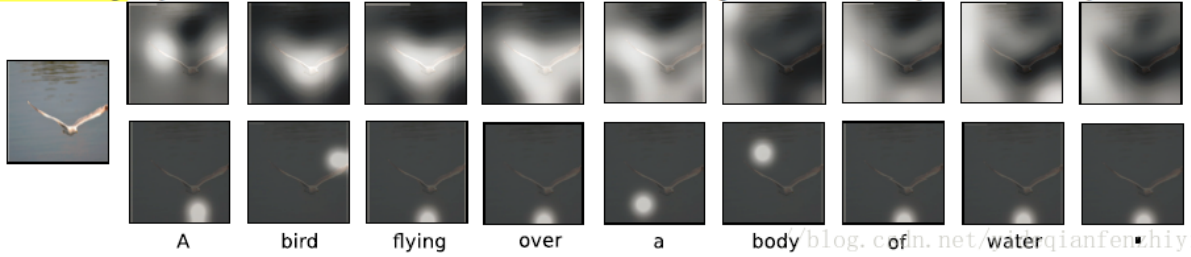


Figure 4. Weight maps produced by max-pooling (row 2) and by attention model (row 3). Notice that our attention model learns better interpretable weight maps for different scales. (a) Scale-1 attention (*i.e.*, weight map for scale $s = 1$) captures small-scale objects, (b) Scale-0.75 attention usually focuses on middle-scale objects, and (c) Scale-0.5 attention emphasizes on background contextual information.

&4.（Image Caption看图说话）Kelvin Xu，etc. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention，ICML2015

因为不做NLP，所以这个论文技术细节并没有看懂。大意是对一个图像进行描述时，生成不同的单词时，其重点关注的图像位置是不同的，可视化效果不错。

Figure 3. Visualization of the attention for each generated word. The rough visualizations obtained by upsampling the attention weights and smoothing. (top)"soft" and (bottom) "hard" attention (note that both models generated the same captions in this example).



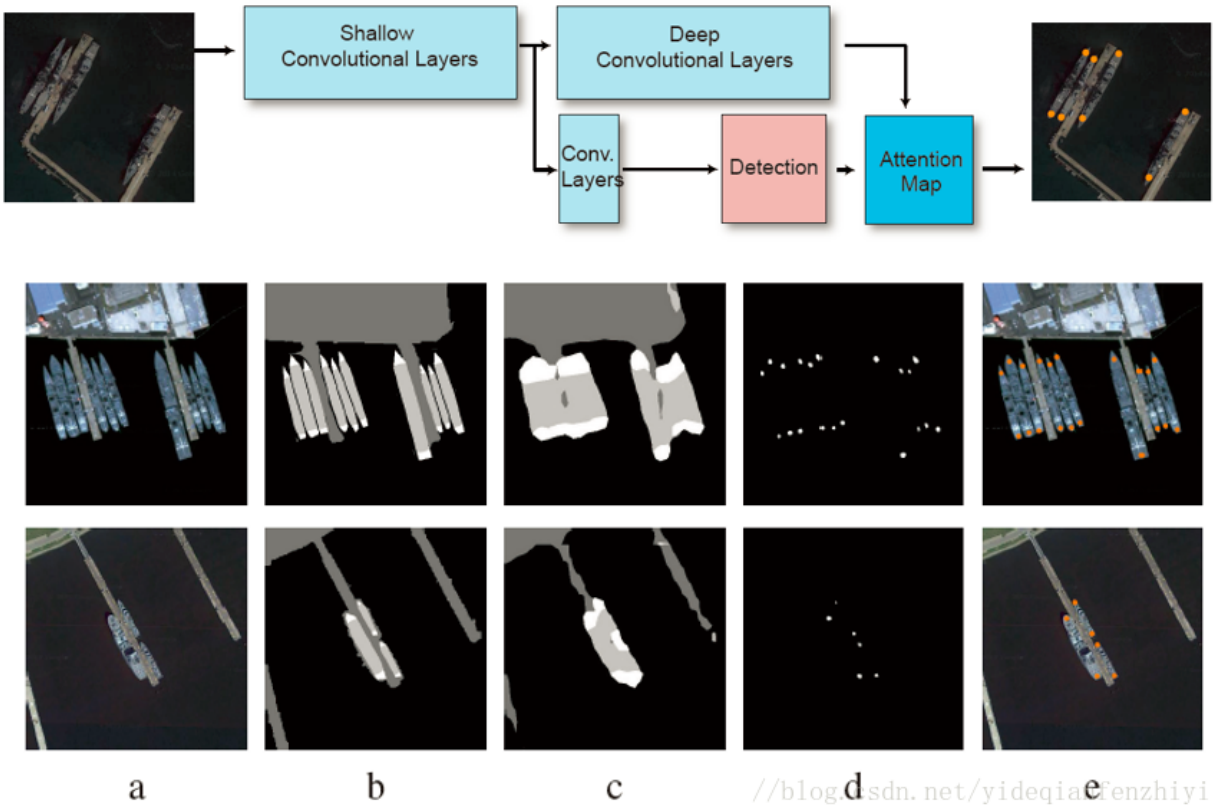
2.2 方式二：任务聚焦/解耦

&1.（Instance Segmentation）Kaiming He, etc. Mask R-CNN（非常好的一篇文章）

Kaiming大神在Mask R-CNN中，将segment branch的损失函数由softmax loss换成了binary sigmoid loss。即是，**将分类和分割任务进行解耦，当box branch已经分好类时，segment branch 就不用再关注类别，只需要关注分割，从而使网络更加容易训练。**具体到训练中，假设分狗、猫、马三类，segment branch会得到3个mask，当训练样本是狗类，那么这个类别的loss才会被反传，猫类和马类对应的mask都不用管。也就是说，**生成狗mask的那部分网络连接（卷积核）只需要聚焦于狗类的样本，然后将属于狗的像素目标凸显出来出来，训练其他类别时不会对这些连接权重进行更新。**通过这个任务解耦，分割的结果得到了很大的提升（5%-7%）。Kaiming大神在文中也指出，当只输出一个mask时，分割结果只是略差，从而进一步说明了将分类和分割解耦的作用。

&2. （图像分割）Lin etc. Fully Convolutional Network with Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images

针对靠岸舰船，本文通过任务解耦的方法来处理。因为高层特征表达能力强，分类更准，但定位不准；底层低位准，但分类不准。为了应对这一问题，本文利用一个深层网络得到一个粗糙的分割结果图（船头/船尾、船身、海洋和陆地分别是一类）即Attention Map；利用一个浅层网络得到船头/船尾预测图，位置比较准，但是有很多虚景。**训练中，使用Attention Map对浅层网络的loss进行引导，只反传在粗的船头/船尾位置上的loss，其他地方的loss不反传。**相当于，**深层的网络能得到一个船头/船尾的大概位置，然后浅层网络只需要关注这些大概位置，然后预测出精细的位置，图像中的其他部分（如船身、海洋和陆地）都不关注，从而降低了学习的难度。**



3. 感想

总的来说，我觉得attention这个概念很有趣，使用attention也可以做出一些有意思的工作。相比于方式一，个人更喜欢方式二任务解耦，因为其对所解决的任务本身有更深刻的认识。当然上述介绍的论文，主要是关于high-level的任务，还没看到attention在low-level的任务中的应用（也可能是自己查得不全），当然如何应用，这值得思考。

参考资料

除了上面的一些论文，其他的参考资料：

知乎问题：目前主流的attention方法都有哪些？

知乎问题：Attention based model 是什么，它解决了什么问题？

知乎专栏总结：计算机视觉中的注意力机制

CSDN博客总结：Attention Model（mechanism） 的 套路

CSDN专栏：从2017年顶会论文看 Attention Model

CSDN专栏：模型汇总24 - 深度学习中Attention Mechanism详细介绍：原理、分类及应用