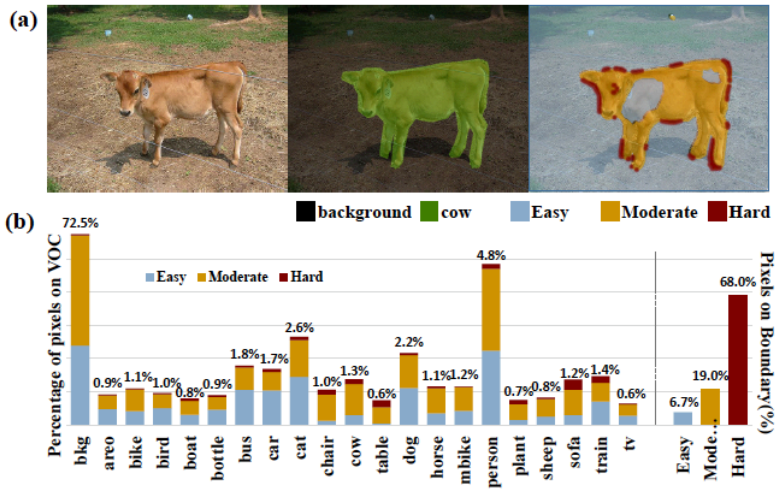
1. 第二篇 CVPR 17 [Not All Pixels Are Equal: Difficulty-Aware Semantic Segmentation via Deep Layer Cascade]
   来自 The Chinese University of Hong Kong

(1) 问题定义



在语义分割问题中，网络最终的输出结果为，对于每个 pixel，都是一个 Nx1 维的向量，N 表示总共有多少个 classes。向量中每个值表示属于对应类别的概率值，通常取最大的概率值表示这个 pixel 属于这个类别。

作者在这里根据每个 pixel 的置信度将所有像素分为三类：

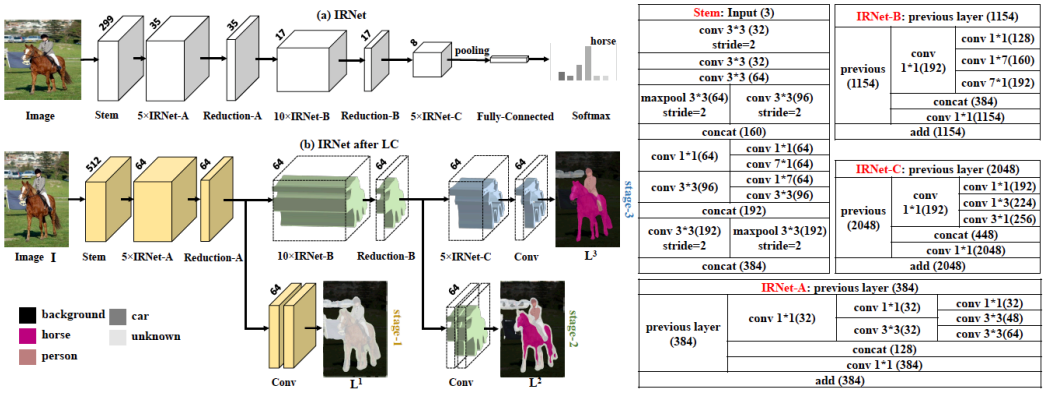
i. easy set (ES)，被正确分类的置信度大于 95%
ii. hard set (HS)，被错误分类的置信度大于 95%
iii. moderate set(MS)，分类概率小于 95%

对 PASCAL VOC 数据集中所有图片进行分析后发现，在每个物体中平均有 30%的像素属于较容易学习的部分，即较低层的网络即可以正确分类每个像素。因此这些像素不需要经过过多层的网络，否则会浪费过多的计算成本。

其次，大约 70%的 HS 中的像素位于物体的边界，因为边界容易有歧义，难以学习。所以对这些像素学习过多可能会导致过拟合。

(2) 提出方法

- 2-1 网络结构



Stem: Input (3)	
conv 3*3 (32) stride=2	
conv 3*3 (32)	
conv 3*3 (64)	
maxpool 3*3(64) stride=2	conv 3*3(96) stride=2
concat (160)	
conv 1*1(64)	conv 1*1(64)
	conv 7*1(64)
conv 3*3(96)	conv 1*7(64)
	conv 3*3(96)
concat (192)	
conv 3*3(192) stride=2	maxpool 3*3(192) stride=2
concat (384)	

IRNet-B: previous layer (1154)		
previous (1154)	conv 1*1(192)	conv 1*1(128)
		conv 1*7(160)
		conv 7*1(192)
	concat (384)	
	conv 1*1(1154)	
	add (1154)	

IRNet-C: previous layer (2048)		
previous (2048)	conv 1*1(192)	conv 1*1(192)
		conv 1*3(224)
		conv 3*1(256)
	concat (448)	
	conv 1*1(2048)	
	add (2048)	

IRNet-A: previous layer (384)			
previous layer (384)	conv 1*1(32)	conv 1*1(32)	conv 1*1(32)
			conv 3*3(48)
		conv 3*3(32)	conv 3*3(64)
	concat (128)		
	conv 1*1 (384)		
	add (384)		

Figure 2: (a) depicts the Inception-ResNet-v2 (IRNet) for classification task. (b) is the architecture of Layer Cascade IRNet (IRNet-LC). The tables at the right show the structure of IRNet.

这里作者提出了级连的三阶段方式。这里以 IRNet（在 PASCAL VOC 12 上比 ResNet 性能高 1.2%）作为骨干网络，其实也可以采用 ResNet 的结构。

以第一阶段为例，在经过 Reduction-A 后，走向了下边分支，经过一个卷积层和 Softmax 层。对每个像素都进行筛选，对每个点其最大分类概率为 Pmax。设置一个 threshold（一般大于 0.95），当 Pmax > threshold 时，即为置信度大于 0.95 的像素，包括 ES 和 HS。这些像素被留在第一阶段进行分类，其余的 MS 像素被传递到下一阶段。第二和第三阶段与此类似。

这样较容易学习的类别如 background 就被留在了靠前的阶段学习，而较难学习的类别可能造成过拟合，也被留在了靠前的阶段。**在避免过拟合的同时也降低了计算成本，使得网络更多的集中注意力在 foreground 上。**

- 2-2 区域卷积

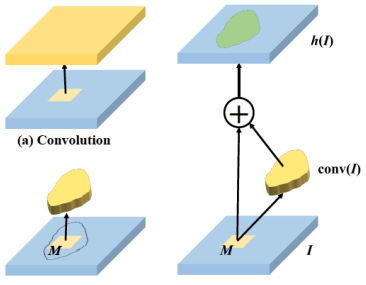


Figure 3: (a) shows the conventional convolution that operates on an entire image. (b) is region convolution (RC) where filters only convolve irregular region of interest denoted as M. Values of the other region are set as zeros. (c) illustrates RC in a residual module. **Best viewed in color.**

由于只有部分像素被传递到下一阶段，因此从第二阶段开始，所进行卷积的图片就是不规则形状的，而非之前的矩形图片。这里作者将不规则形状之外的位置全部填充为 0 然后进行卷积运算。

(3) 实验结果

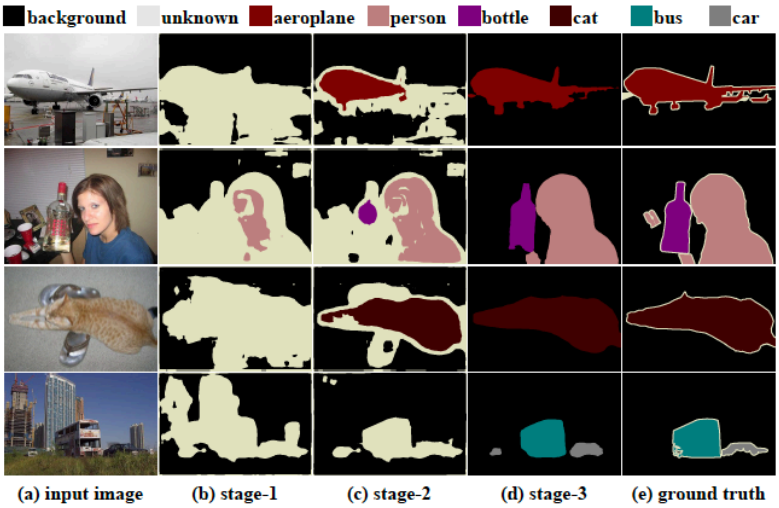


Figure 4: Visualization of different stages' outputs in VOC12 dataset. **Best viewed in color.**

一个可视化结果。靠后的阶段会有更好的性能，但是注意由于将边缘的分类抛给了第一阶段，导致边缘像素没有被很好的学习，因此边缘部分的判断并不够准确，如 d 列图 2 的酒瓶下端。

(4) 结论与思考

- 优点

这篇文章最大的意义是利用很好的可视化方法揭示了不同像素的学习难度差异。

另一个具有启发性的地方在于 stage-2 的学习材料更少了，也可以认为是噪声更少了，模型能够更加专注于中等学习难度的事情了（不允许把对占有极小部分的 HS 送入更深层的网络进行学习）。而 stage-3 相对于 stage-2 来说，更加专注于第二阶段中的 MS 难度的区域。**这些区域的置信度不够高，说明需要更多的学习，来增加对其的判断能力。**

- 缺点

但是在方法上的贡献更多的在于计算成本的降低，而非分割性能的提升。

实验结果中的多阶段示意图，很大部分原因是随着网络的加深，有更高层的语义信息被学习到。即使不加上 LC 模块，而直接可视化不同层的分类结果，也会有类似这样的效果。

- 思考

对于置信度不够高的区域，不容易被识别出其所属种类，这种区域可能需要设计专门的网络进行学习。本文提出的结构相当于对于较难给出正确分类（而非较容易给出错误分类的 HS）的 MS 进行更深层次的学习。

但是也可以设计专门的网络结构来分类 MS 区域。MS 区域的难分类，可能是由于尺寸变化等多种原因造成的