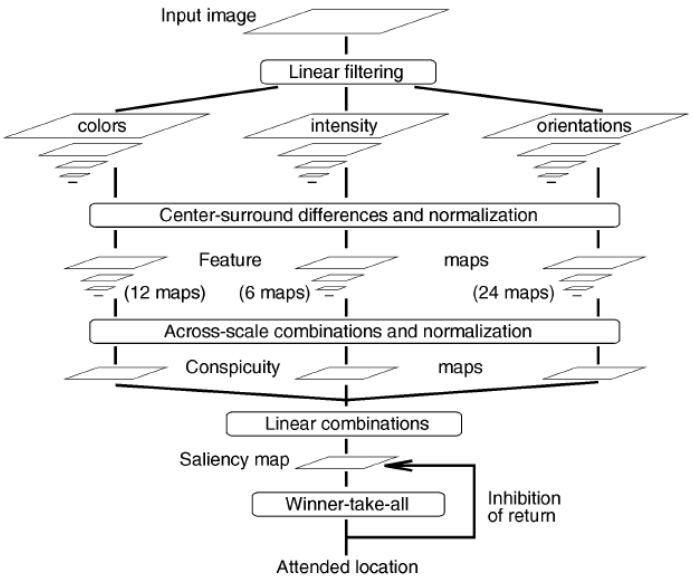
注意力机制 Review

1 发展

原始版本：Itti L. [1]，一篇关于图像显著性的文章，引用量：9843。

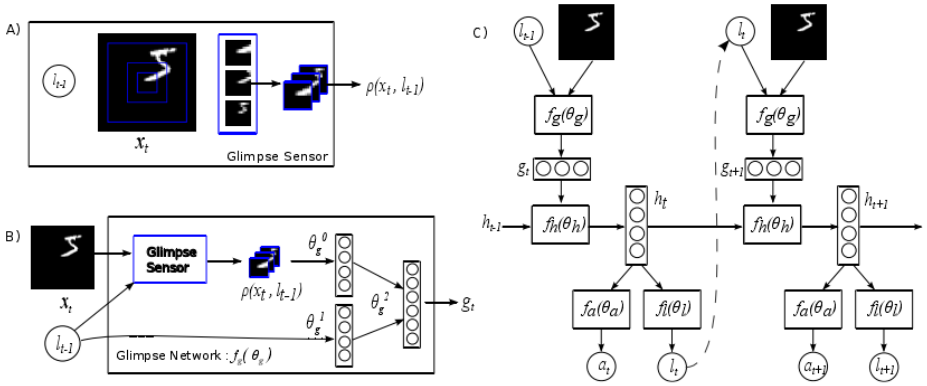


该模型把输入图像进行早期视觉抽样，分解成三个通道：强度，颜色（R、B、G、Y）和方向（0°，45°，90°，135°）。通过中心-周围算子，即不同尺度的滤波器响应差值来产生一组特征图。然后把每个通道的特征图归一化，并跨尺度和方向相结合，产生每个通道的突出图。然后这些通道线性组合形成整体的显著图。这个模型已经证明在预测人的注意力和目标检测上是很有用的。

注：在高斯金字塔中，尺度较大的图像细节信息较多，而尺度较小的图像由于高斯平滑和减抽样操作使得其更能反映出局部的图像背景信息，因而将尺度较大的图像和尺度较小的图像进行跨尺度减操作（across-scale），能得到局部中心和周边背景信息的反差信息。跨尺度减的具体算法如下：通过将代表周边背景信息的较小尺度的图像进行线性插值，使之与代表中心信息的较大尺度的图像具有相同大小，然后进行点对点的减操作。

转折点：

图像：google mind 团队的论文[2]，引用量：783。他们在 RNN 模型上使用了 attention 机制来对手写数据集进行分类。



NLP：随后，Bahdanau 等人在论文[3]中，使用类似 attention 的机制在机器翻译任务上将翻译和对齐同时进行，他们的工作算是是第一个提出 attention 机制应用到 NLP 领域中。引用量：4718。

2 Soft Attention 和 Hard Attention

Kelvin Xu 等人与2015年发表论文[4]，在 Image Caption 中引入了 Attention，当生成第 i 个关于图片内容描述的词时，用 Attention 来关联与 i 个词相关的图片的区域。在论文中使用了两种 Attention Mechanism，即 Soft Attention 和 Hard Attention。传统的 Attention Mechanism 就是 Soft Attention。Soft Attention 是参数化的（Parameterization），因此可导，可以被嵌入到模型中去，直接训练。梯度可以经过 Attention Mechanism 模块，反向传播到模型其他部分。

相反，Hard Attention 是一个随机的过程。Hard Attention 不会选择整个 encoder 的输出做为其输入，Hard Attention 会依概率 Si 来采样输入端的隐状态一部分来进行计算，而不是整个 encoder 的隐状态。为了实现梯度的反向传播，需要采用蒙特卡洛采样的方法来估计模块的梯度。

两种 Attention Mechanism 都有各自的优势，但目前更多的研究和应用还是更倾向于使用 Soft Attention，因为其可以直接求导，进行梯度反向传播。

3 应用

NLP：(待完善)

https://zhuanlan.zhihu.com/p/31547842

https://www.cnblogs.com/robert-dlut/p/8638283.html

图像：

重要分支：

1) fixation prediction，旨在预测出图像中的注视点，这个注视点有可能是 bottom-up 与任务无关的，还有可能是 top-down 与当前任务相关的；
2) salient object detection（显著性物体检测）；
3) objectness proposals，它是基于窗口的度量方法，通过预测图像中的每个窗口有多大可能性含有物体，有助于后期做物体检测。

“注意力”的方式：

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。

   - 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），此时常用强化学习来做。

   - 这个加权可以作用在原图上，也就是[2]（RAM）和[5]（DRAM）；也可以作用在特征图上，如后续的文章（例如 image caption 中的[4])。

   - 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。

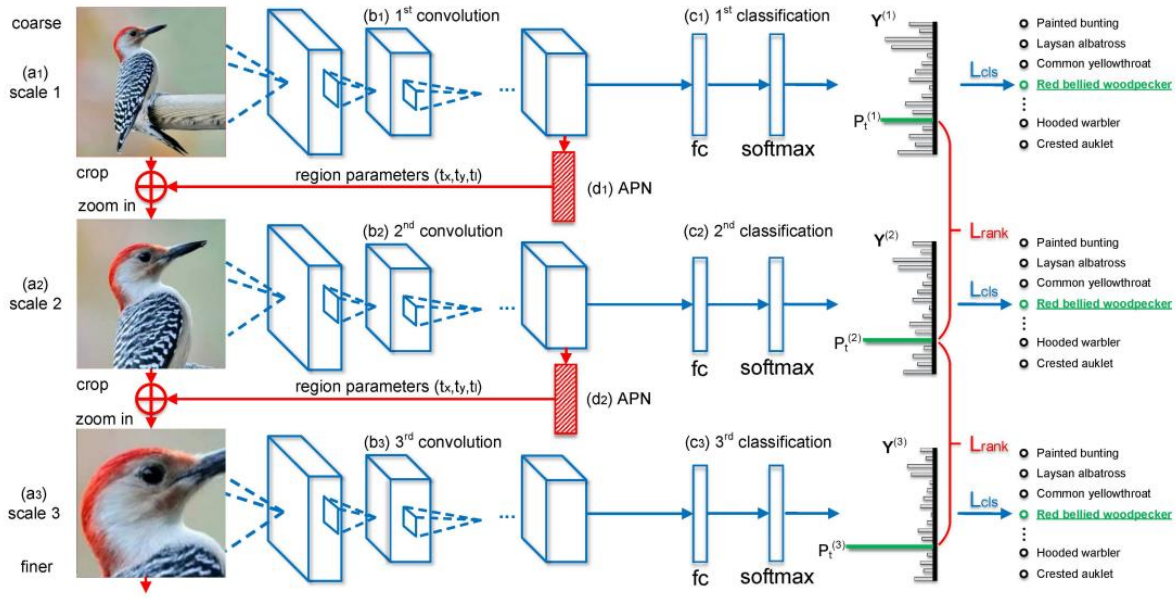
2) 任务聚焦：通过将任务分解，设计不同的网络结构（或分支）专注于不同的子任务，重新分配网络的学习能力，从而降低原始任务的难度，使网络更加容易训练。

应用点：

1) 精细分类

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

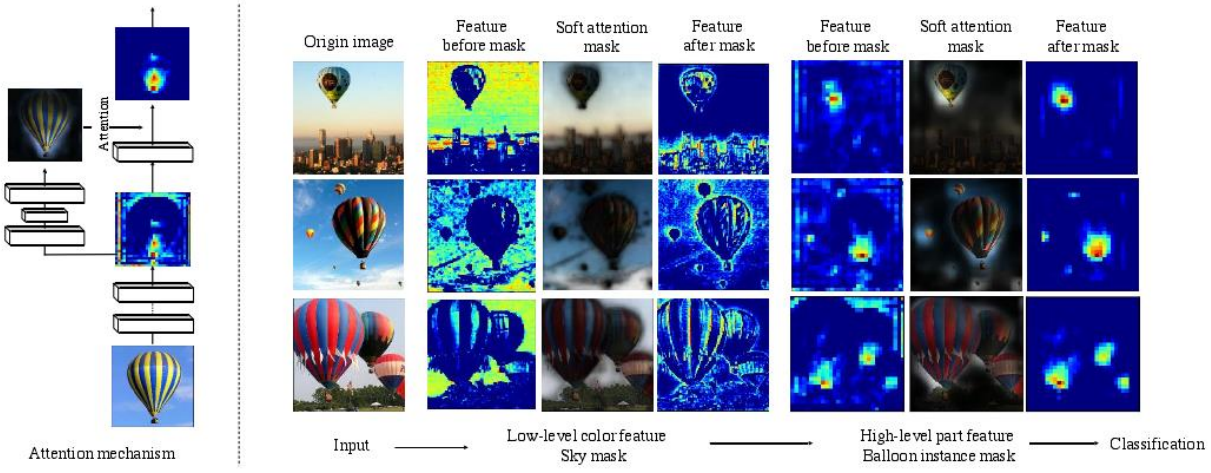
  在关注的每一个目标尺度上，都采用一个分类的网络和一个产生 attention proposal 的网络(APN)。本文最有趣的就是这个 APN。这个 APN 由两个全连接层构成，输出3个参数表示方框的位置，接下来的尺度的分类网络只在这个新产生的方框图像中提特征进行分类。怎么训练呢？本文定义了一个叫做 rank Loss，用这个 loss 来训练 APN，并强迫 finer 的尺度得到的分类结果要比上一个尺度的好，从而使 APN 更提取出更有利于精细分类的目标局部出来。通过交替迭代训练，APN 将越来越聚焦目标上的细微的有区分性的部分。



2) 图像分类

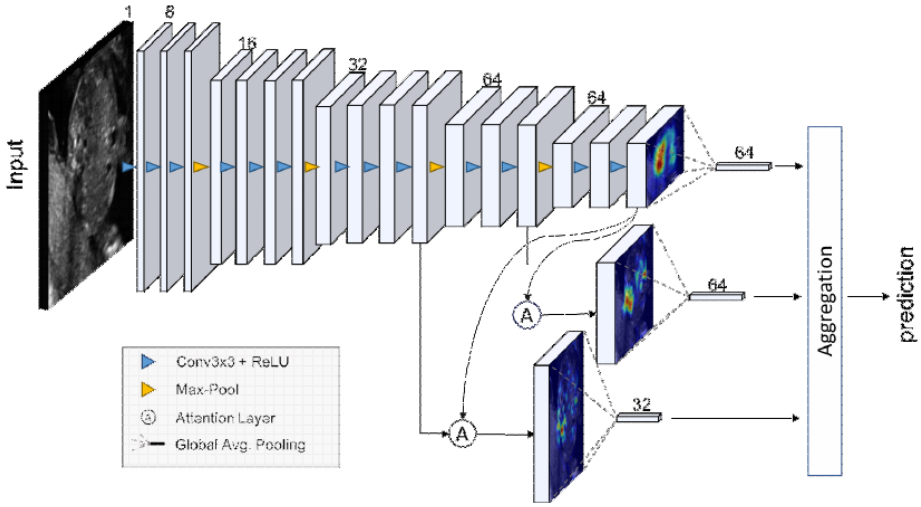
- [7] Residual Attention Network for Image Classification，CVPR2017.

  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。



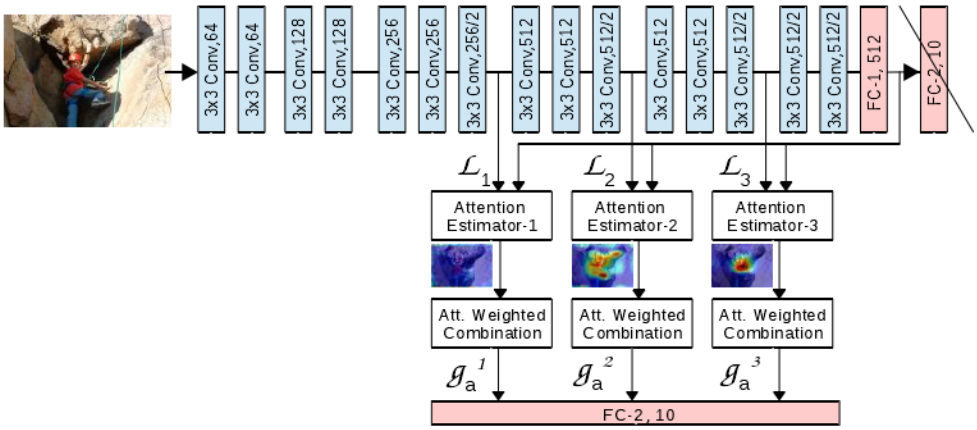
- [16] Attention-Gated Networks for Improving Ultrasound Scan Plane Detection

  建议引入 self-gated 软注意机制，产生一个端到端可训练的模型，它允许网络有效地利用对预测有用的上下文局部信息。在 U-Net 基础上



- [14] Learn To Pay Attention

  提出了一种用于卷积神经网络结构的端到端可训练注意模块，用于图像分类。该模块将二维特征向量映射作为输入图像在 CNN 管道中不同阶段的中间表示，并为每个映射输出一个二维矩阵分数，学习过的注意力地图在抑制背景混乱的同时，突出了感兴趣的区域。得到一张与 feature map 分辨率相同的 score map, 再将这张 score map 做归一化，就得到了 attention map。以 attention map：[a1, a2, a3, ...an]为权重，对 local feature map 的所有 feature(li) 做加权求和: a1x l1 + a2 x l2 +... an x ln. 得到的 feature 称为 attention feature, 作为用于最终分类的特征。



- [15] Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer（还没看懂）

  将注意力模型用于 Teacher-Student 蒸馏过程。文中提到把 teacher 模型中 loss 对 input 的导数作为知识传递给 student 模型，因为 loss 对 input 的导数反映了网络 output 的变化对于 input 的敏感程度，如果某个像素一个小的变化对于网络输出有一个大的影响，我们就可以认为网络"pay attention"那个像素。

3) Image Caption 看图说话

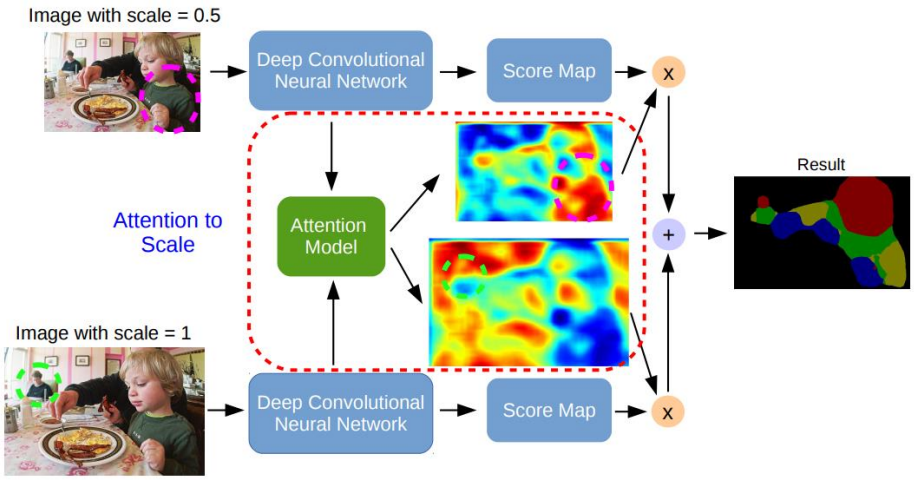
- [4] Show, Attend and Tell: Neural Image Caption Generation with Visual Attention，ICML2015

  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。

4) 图像语义分割

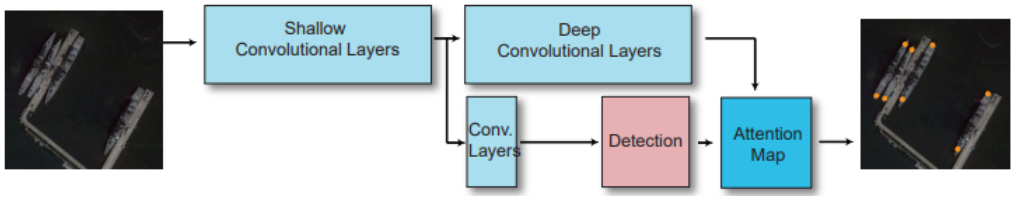
- [8] Attention to Scale: Scale-aware Semantic Image Segmentation, CVPPR2016

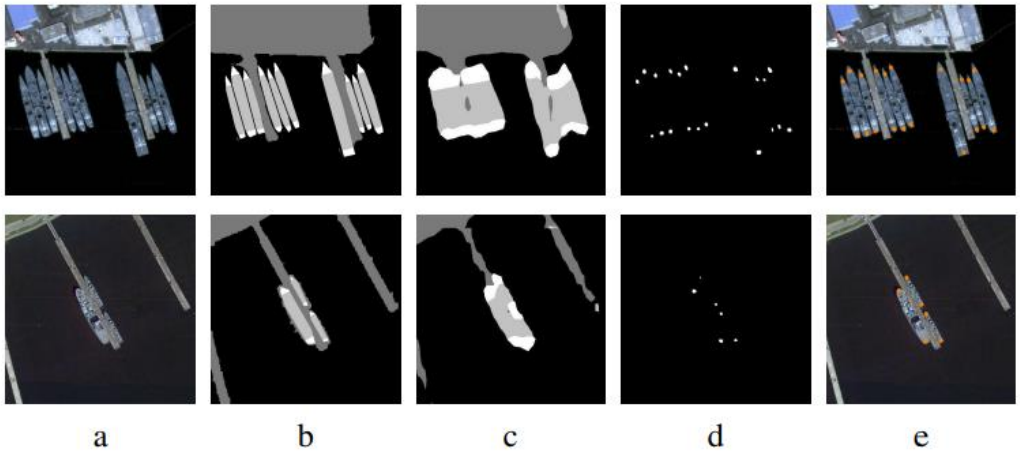
通过对输入图片的尺度进行放缩，构造多尺度。传统的方法是使用 average-pooling 或 max-pooling 对不同尺度的特征进行融合，而本文通过构造 Attention model（由两个卷积层构成）从而自动地去学不同尺度的权重，进行融合（效果提升1到2个点吧，不同的数据集不一样）。从论文中的权重可视化的结果，能发现大尺寸输入上，对应网络关注于 small-scale objects，而在稍微小一点的尺寸输入上，网络就关注于 middle-scale，小尺寸输入则关注 background contextual information。



- [9] Fully Convolutional Network with Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images

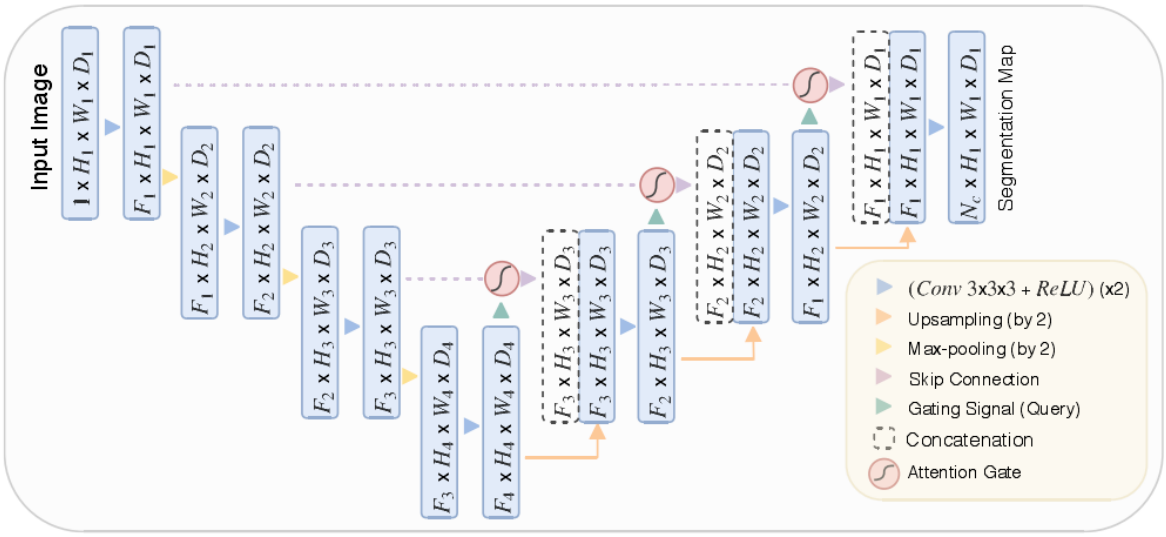
针对靠岸舰船，本文通过任务解耦的方法来处理。因为高层特征表达能力强，分类更准，但定位不准；底层低位准，但分类不准。为了应对这一问题，本文利用一个深层网络得到一个粗糙的分割结果图（船头/船尾、船身、海洋和陆地分别是一类）即 Attention Map；利用一个浅层网络得到船头/船尾预测图，位置比较准，但是有很多虚景。训练中，使用 Attention Map 对浅层网络的 loss 进行引导，只反传在粗的船头/船尾位置上的 loss，其他地方的 loss 不反传。相当于，深层的网络能得到一个船头/船尾的大概位置，然后浅层网络只需要关注这些大概位置，然后预测出精细的位置，图像中的其他部分（如船身、海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





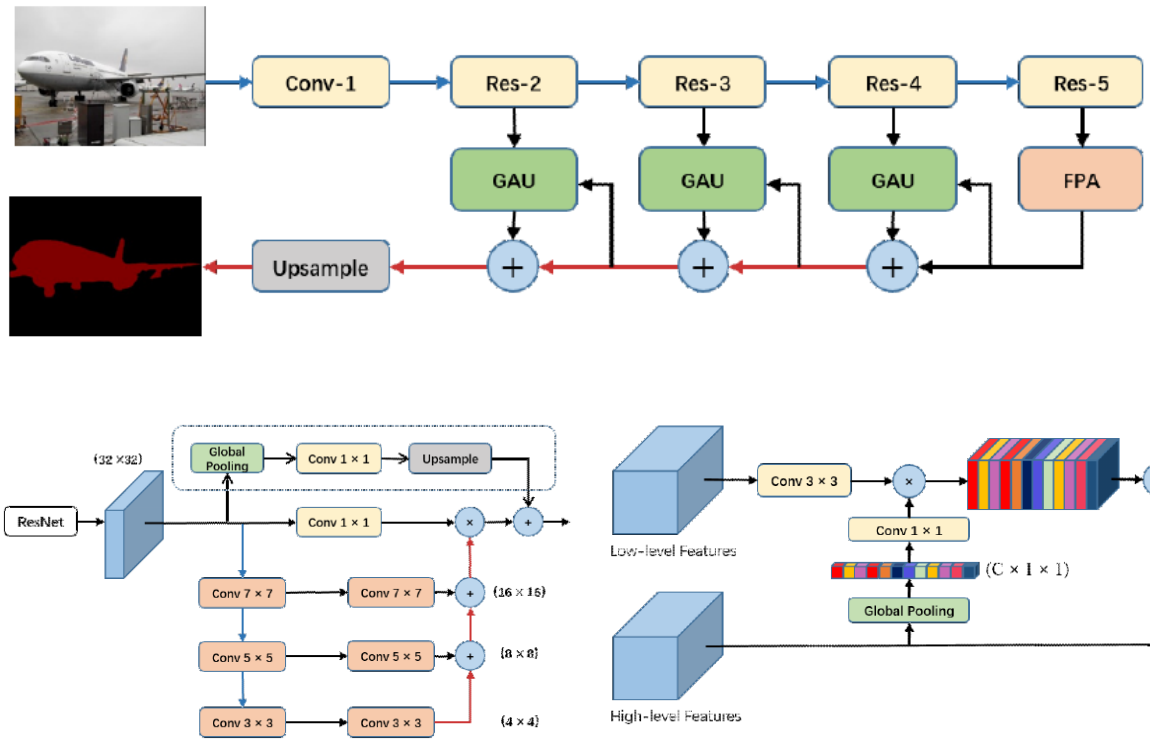
- [10] Attention U-Net: Learning Where to Look for the Pancreas. 2018.

  提出了一种用于医学成像的新型注意门模型，该模型自动学习聚焦于不同形状和大小的目标结构。在 U-Net 基础上加上权重注意力机制，优化分割效果。用 AG 训练的模型隐含地学习抑制输入图像中的不相关区域，同时突出显示对特定任务有用的显着特征。



- [11] Li H, Xiong P, An J, et al. Pyramid Attention Network for Semantic Segmentation[J]. 2018

  引入了一个特征金字塔注意力模块（Feature Pyramid Attention module），在高层的输出上施加空间金字塔注意力结构，并结合全局池化策略来学习更好的特征表征。此外，利用每个解码器层中的全局注意力上采样模块（Global Attention Upsample module）得到的全局上下文特征信息，作为低级别特征的指导，以此来筛选不同类别的定位细节。论文作者表示，他们提出的方法在 PASCAL VOC 2012 数据集上实现了当前最佳的性能。而且无需经过 COCO 数据集的预训练过程，他们的模型在 PASCAL VOC 2012 和 Cityscapes 基准测试中能够实现了 84.0% mIoU。

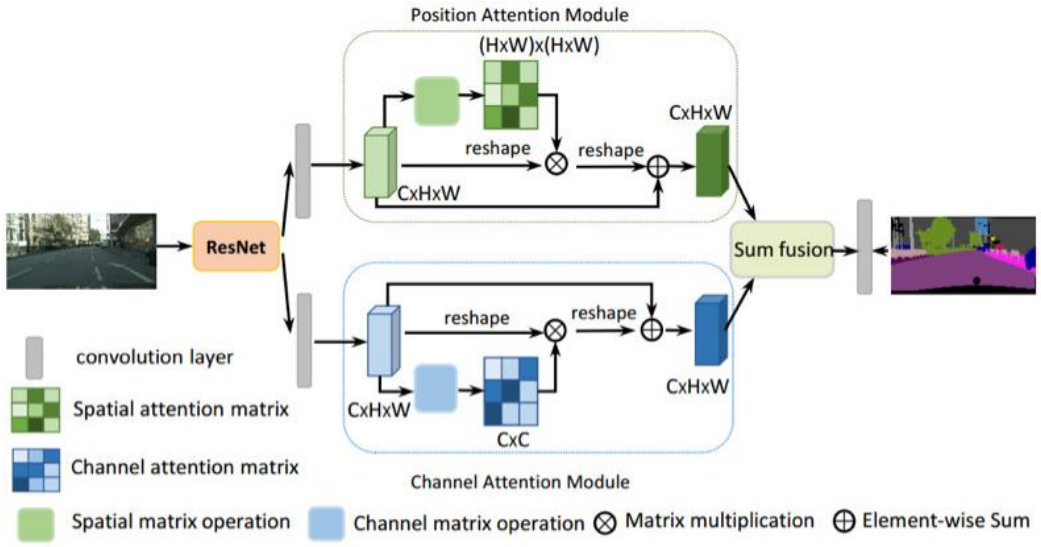


(b) Feature Pyramid Attention

Figure 4: Global Attention Upsample module structure

- [12] Dual Attention Network for Scene Segmentation. 2018.

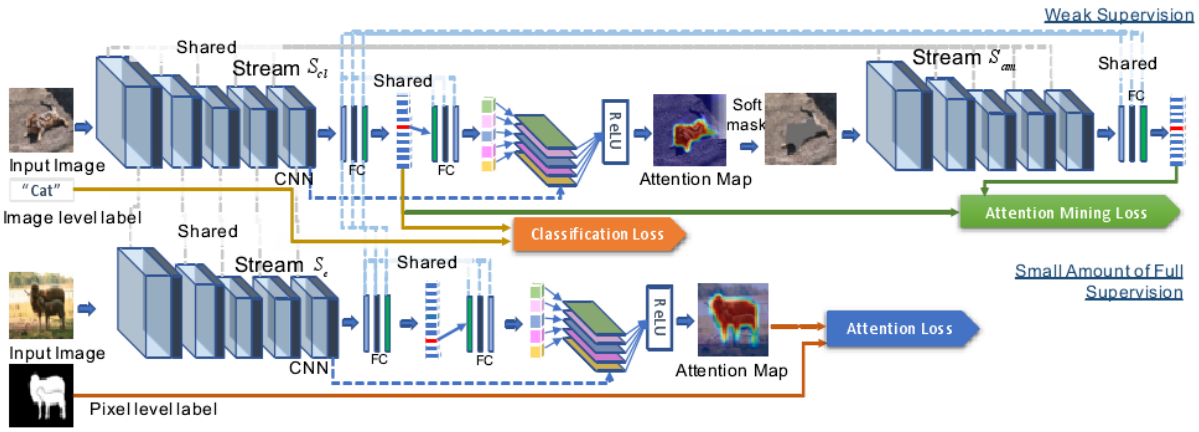
  提出了一种简单有效的双重注意力网络（Dual Attention Network, DANet），通过引入自注意力机制 (self-attention mechanism) 在特征的空间维度和通道维度分别抓取特征之间的全局依赖关系，增强特征的表达能力。该网络在 Cityscapes，PASCAL Context 和 COCO-Stuff 三个公开的场景分割数据集上均取得了当前最好性能，相比 Dilated FCN 性能得到 5 个点以上的显著提升。



5) 图像弱监督分割

- [13] Tell Me Where to Look: Guided Attention Inference Network. 2018.

  网络主要分为两个部分，第一部分为分类部分，就是简单的使用 VGG 对图像进行特征提取，然后使用全链接层实现对物体的识别。所以本论文的主要创新点主要集中在第二部分上。第二部分为注意力提取部分，它在很大程度上引用了 Grad-CAM 的思想，通过预测层对最后一层的 feature_map 求导之后再对其做 globalaveragepool(GAP)。但是!! 他加入了样本中5%-10%的像素级标签，不应该完全算作是弱监督了。



6) 预测

- [17] Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks

  介绍了单个 Attention Model 在医疗诊断预测中的应用。这个模型的输入是用户前 t 次的医疗诊断结果（用 one-hot 的形式表示，如果结果中存在某一医疗代码则向量对应位置上值为1，否则为0），输出是用户下一时刻的医疗诊断类型。使用 Attention Model 的思想是：用户下一时刻被诊断的疾病类型可能更与前面某一次或某几次的医疗诊断相关。

文献列表

[1] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[M]. IEEE Computer Society, 1998.

[2] Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention[J]. 2014, 3:2204-2212.

[3] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate[J]. Computer Science, 2014.

[4] Xu K, Ba J, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[J]. Computer Science, 2015:2048-2057.

[5] Ba J, Mnih V, Kavukcuoglu K. Multiple object recognition with visual attention[J]. arXiv preprint arXiv:1412.7755, 2014.

[6] Fu J, Zheng H, Mei T. Look Closer to See Better: Recurrent Attention Convolutional Neural Network for Fine-Grained Image Recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2017:4476-4484.

[7] Wang F, Jiang M, Qian C, et al. Residual Attention Network for Image Classification[C]// Computer Vision and Pattern Recognition. IEEE, 2017:6450-6458.

[8] Chen L C, Yang Y, Wang J, et al. Attention to Scale: Scale-Aware Semantic Image Segmentation[C]// Computer Vision and Pattern Recognition. IEEE, 2016:3640-3649.

[9] Lin H, Shi Z, Zou Z. Fully Convolutional Network With Task Partitioning for Inshore Ship Detection in Optical Remote Sensing Images[J]. IEEE Geoscience & Remote Sensing Letters, 2017, PP(99):1-5.

[10] Oktay O, Schlemper J, Folgoc L L, et al. Attention U-Net: Learning Where to Look for the Pancreas[J]. 2018.

[11] Li H, Xiong P, An J, et al. Pyramid Attention Network for Semantic Segmentation[J]. 2018.

[12] Fu J, Liu J, Tian H, et al. Dual Attention Network for Scene Segmentation[J]. 2018.

[13] Li K, Wu Z, Peng K C, et al. Tell Me Where to Look: Guided Attention Inference Network[J]. 2018.

[14] Jetley S, Lord N A, Lee N, et al. Learn To Pay Attention[J]. 2018

[15] Zagoruyko S, Komodakis N. Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer[J]. 2016.

[16] Schlemper J, Oktay O, Chen L, et al. Attention-Gated Networks for Improving Ultrasound Scan Plane Detection[J]. 2018.

[17] Ma F, Chitta R, Zhou J, et al. Dipole: Diagnosis Prediction in Healthcare via Attention-based Bidirectional Recurrent Neural Networks[J]. 2017:1903-1911.