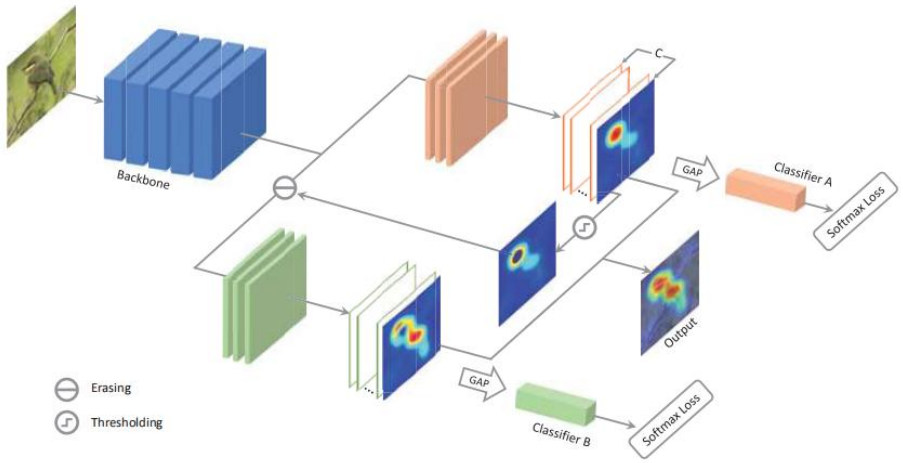
论文学习 1：Adversarial Complementary Learning for weakly supervised object localization
这篇能应用到当前的框架比较多

主要思想：对于一张给定输入，由于 CAM 会出现背景激活过度或者前景激活不足的情况，构造两个独立网络（记为 A、B）对其进行特征提取，当 A 提取完特征后，在 B 提取特征时把 A 的特征进行擦除，以强迫 B 网络对未激活的特征部分进行学习，最后将两个网络的特征图输出融合在一起。

缺点：需要训练多个独立网络来获得对象区域，势必会花费更多的训练时间和计算资源；在没有别的监督信息指导下，网络并不一定总能发现新的对象区域。



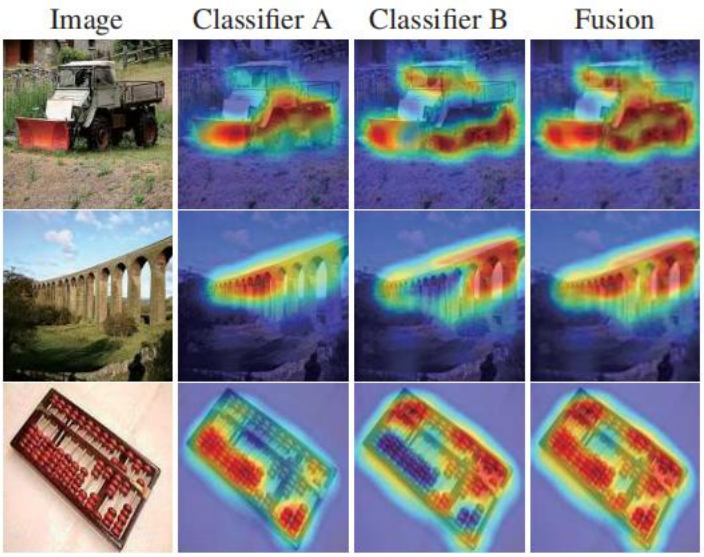
基本架构

简要说明：Backbone 是一个全卷积网络，用于特征提取，然后接两个分类器，其输入特征不同，B 的特征输入会在原输入特征上擦除 A 的定位映射特征。具体是，先对 A 的特征定位映射加以阈值判断，将大于阈值（认为特征明显）的部分设为 0，以对抗的方式擦除 B 的输入部分，从而激励 B 学习到其他的 A 未关注的特征，最后将两个分支结合起来获得完整定位区域。

Algorithm 1 Training algorithm for ACoL

Input: Training data $I = \{(I_i, y_i)\}_{i=1}^{N}$, threshold δ
while training is not convergent **do**
  Update feature maps $S \leftarrow f(\theta_0, I_i)$
  Extract localization map $M^A \leftarrow f(\theta_A, S, y_i)$
  Discover the discriminative region $R = \bar{M}^A > \delta$
  Obtain erased feature maps $\tilde{S} \leftarrow erase(S, R)$
  Extract localization map $M^B \leftarrow f(\theta_A, S, y_i)$
  Obtain fused map $\bar{M}_{i,j}^{fuse} = \max(\bar{M}_{i,j}^{A}, \bar{M}_{i,j}^{A})$
  Update θ_0, θ_A and θ_B
end while
Output: $\bar{M}^{fuse}$

结果展示：



一些新的方法：

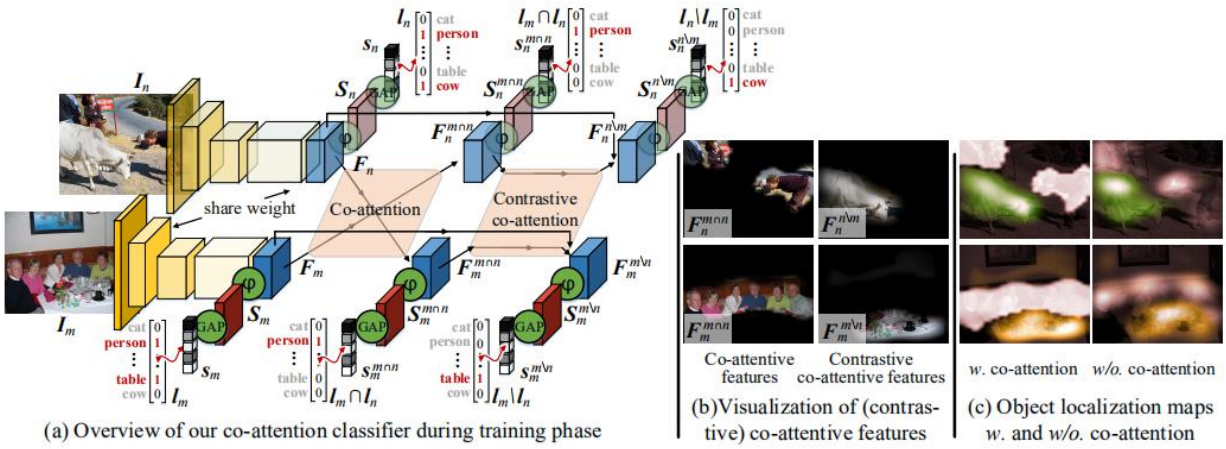
CAM 新生成思路：之前的 CAM 是通过某层特征图和网络输出之间求梯度等一些运算后得到，且不能直接做端到端的训练（图片输入麻烦）。这篇文章从数学角度证明了可以通过一步步骤得到与 CAM 效果极为相似的图。

Given the output feature maps S of an FCN, we add a convolutional layer of C channels with the kernel size of 1 × 1, stride 1 on top of the feature maps S. Then, the output is fed into a GAP layer followed by a softmax layer for classification

通过这种修正方法，对象定位的映射就可以直接在前向传递中获得，而不需要 CAM 的后处理步骤。

作者处理结果对比图

论文学习 2：Mining Cross-Image Semantics for Weakly Supervised Semantic Segmentation
来自 2020CVPR



(a) Overview of our co-attention classifier during training phase
(b)Visualization of (contrastive) co-attentive features
(c) Object localization maps w. and w/o. co-attention

基本框架

思想概述：第一部分依然是特征提取，输入位两张图片 m，n，输出特征图记为 Fm 和 Fn，随后引入两个注意力机制，其中一个是同类别的注意力机制，另一个是异类别的注意力机制，在这篇文章里，输入的两张图拥有同样的类别（比如 person）和不同的类别（比如 m 中有 cow，而 n 中有 table），两个注意力机制是用于寻找相同的部分和不同的部分。

首先是同类别注意力部分：

their correlations. We first compute the affinity matrix $\boldsymbol{P}$ between $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$:

$$\boldsymbol{P} = \boldsymbol{F}_m^{\top} \boldsymbol{W}_{\boldsymbol{P}} \boldsymbol{F}_n \in \mathbb{R}^{HW \times HW}, \tag{2}$$

where $\boldsymbol{F}_m \in \mathbb{R}^{C \times HW}$ and $\boldsymbol{F}_n \in \mathbb{R}^{C \times HW}$ are flattened into matrix formats, and $\boldsymbol{W}_{\boldsymbol{P}} \in \mathbb{R}^{C \times C}$ is a learnable matrix. The affinity matrix $\boldsymbol{P}$ stores similarity scores corresponding to all pairs of positions in $\boldsymbol{F}_m$ and $\boldsymbol{F}_n$, *i.e.*, the $(i, j)^{th}$ element of $\boldsymbol{P}$ gives the similarity between i^{th} location in $\boldsymbol{F}_m$ and j^{th} location in $\boldsymbol{F}_n$.

Then $\boldsymbol{P}$ is normalized column-wise to derive attention maps across $\boldsymbol{F}_m$ for each position in $\boldsymbol{F}_n$, and row-wise to derive attention maps across $\boldsymbol{F}_n$ for each position in $\boldsymbol{F}_m$:

$$\boldsymbol{A}_m = \text{softmax}(\boldsymbol{P}) \in [0, 1]^{HW \times HW}, \quad \boldsymbol{A}_n = \text{softmax}(\boldsymbol{P}^{\top}) \in [0, 1]^{HW \times HW}, \tag{3}$$

$$\boldsymbol{F}_m^{m \cap n} = \boldsymbol{F}_n \boldsymbol{A}_n \in \mathbb{R}^{C \times H \times W}, \quad \boldsymbol{F}_n^{m \cap n} = \boldsymbol{F}_m \boldsymbol{A}_m \in \mathbb{R}^{C \times H \times W}, \tag{4}$$

原文如上，大概是首先将 Fm 和 Fn 展平，通过一个可学习的矩阵 Wp 与 Fm 和 Fn 作乘操作得到相似性矩阵 P，P 中的第(i,j)元素表示了 Fm 和 Fn 在第(i,j)位置的相似度大小，然后在一些归一化等操作后，将得到的相似度矩阵分别乘以原来的特征图，这样做就把相似的部分强化了而把不相似的部分弱化了。

可视化后的样子，只有人的部分显示了。

异类别的操作原理也差不多。

想法：这些是基于一张图中有多个实物标签做的（像是有 person、table、cow），因此它的分割难度可能更大。对于肿瘤 MRI 图，一般只具有背景、肿瘤和头部（骨骼那种框架），而在之前观察健康样本的 CAM 时，也确实骨骼的权值较高，疾病样本出问题的也有在骨骼位置权值高的。有没有必要引入多标签描述一份样本（头颅+肿瘤，头颅+无肿瘤），去弱化掉同样标签的头颅部分？这么做和直接判别两张图是否属于同一类会不会有区别？还有没有其他可以延伸的地方呢？