论文名称：Self-EMD: Self-Supervised Object Detection without ImageNet

论文地址：https://arxiv.org/abs/2011.13677

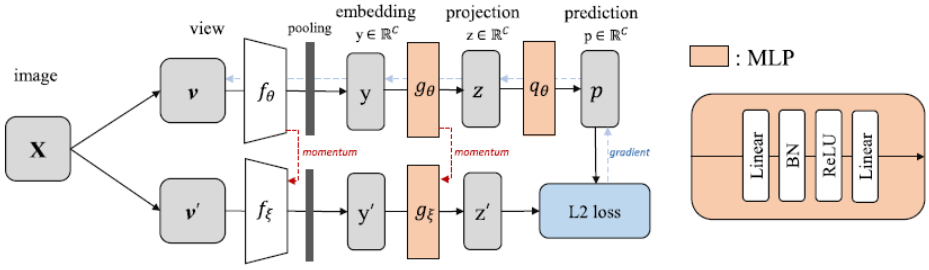
核心思想

提出了一个应用于**目标检测的自监督表示学习方法——self-EMD**，可以直接采用COCO数据集（non-iconic）进行训练，不像传统的方法在ImageNet数据集(iconic-object)上进行训练。利用**卷积特征图作为image embedding**（一般的自监督学习的方法采用经过感知机后的一位向量作为embedding，损失了空间结构，但是目标检测住主要依赖于空间结构），并使用**EMD(Earth Mover's Distance)**来计算一对embedding之间的相似性，最终使用Faster RCNN（ResNet50_FPN)的算法在COCO数据集上39.8%的mAP，与现有的自监督学习的算法（在ImageNet上预训练之后）精度基本上一致，如果采用更多的无标注数据，那么这个算法在COCO上的mAP可以达到40.4%.

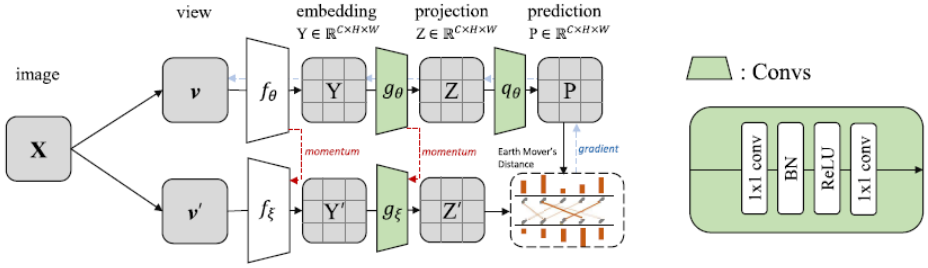
算法介绍

现在的自监督学习通过在ImageNet数据集上实现实例级的图像分类来进行无监督的预训练，通过**最大化不同图像之间的距离（相似度），最小化同一张图像的不同view之间的相似度来学习一个良好的表示**，这种方法针对ImageNet这种分类数据集（一张图像上一个分类物体）来说是适用的。但是针对COCO这种多目标数据集来说就不太适用，因为如果对**一张图像是实现裁剪可能得到的是不同的物体**，因此在目标检测中这种对比表示学习的方法不适用。
而且传统的自监督表示学习的方法常常使用**Global Pooling**的方法来得到Image Embedding，这就**损失了图像的局部与空间信息**，在目标检测中图像的不同位置对应了不同的物体，空间结构比较重要。**self-EMD提出去掉全局池化层**，直接使用**卷积特征图作为Image Embedding**，这样就能保存局部与空间信息，但是这样该如何度量两个feature map之间的相似性呢？而且同一个image的不同crop图像可能包含着不同的图像，因此，度量标准就需要在不同的局部patch中能够挑选出最优的匹配并且最小化不相关区域之间的噪声问题，本文提出使**用EMD（Earth Movier's Distance）**来作为度量标准计算所有局部patches的相似性，这种方法命名为Self-EMD。EMD适用于度量结构性表示之间相似性。给定所有元素对之间的相似性，EMD可以在拥有最小损失的结构之间获得最优的匹配。文章采用cosine相似度来度量两个feature map之间的不同位置之间的相似性并且为EMD约束设置一个合适的权重。

Self-EMD采用BYOL作为其baseline。



(a) BYOL



(b) *Self*-EMD

Self-EMD与BYOL不同在于去掉了最后的全局池化层，并采用卷积层替代了MLP head。采用最后的卷积特征图作为image embedding。

Earth Mover's Distance用来度量两组加权的obejct或者加权的分布之间的距离。离散版本的EMD已经在最优传输问题（OTP）中已经被广泛的研究。特别地，如果需要运输一组资源 $\mathcal{S} = \{s_i | i = 1, 2, ...m\}$ 到一些目的地 $\mathcal{D} = \{d_j | j = 1, 2, ...n\}$，从si到dj的运输损失记为cij， 策略使用 π_{ij}. 来进行标记。最终找寻最优的策略：

$$\min_{\pi} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\pi_{ij}$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \pi_{ij} = s_i, \quad i = 1, 2, ...m,$$

$$\sum_{j=1}^{n} \pi_{ij} = d_j, \quad j = 1, 2, ...n,$$

$$\pi_{ij} \geq 0, \quad i = 1, 2, ...m, j = 1, 2, ...n.$$

线性最优问题，可以在多项式中时间进行求解，但是针对图像特征图，时间复杂度存在图像的分辨率的平方还有batch size，时间复杂度依然很高。使用快速的迭代法（Sinkhorn-Knopp算法）来求解：

$$\min_{\pi} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\pi_{ij} + \gamma E(\pi_{ij}),$$

E为正则化项， where $E(\pi_{ij}) = \pi_{ij}(\log \pi_{ij} - 1)$ and γ is a constant hyper-parameter that controls the intensity of regularization term.

利用拉格朗日变换为无约束的最优问题：

$$\min_{\pi} \quad \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\pi_{ij} + \gamma E(\pi_{ij}) + \alpha_j(\sum_{i=1}^{m}\pi_{ij} - d_j) + \beta_i(\sum_{j=1}^{n}\pi_{ij} - s_i),$$

令导数为0，得到：

$$\pi_{ij}^{*} = \exp(-\frac{\alpha_j}{\gamma})\exp(-\frac{c_{ij}}{\gamma})\exp(-\frac{\beta_i}{\gamma}) \tag{6}$$

Letting $u_j = \exp(-\frac{\alpha_j}{\gamma}), v_i = \exp(-\frac{\beta_i}{\gamma}), M_{ij} = \exp(-\frac{c_{ij}}{\gamma})$, we then enforce the constraints:

$$\sum_i \pi_{ij} = u_j(\sum_i M_{ij}v_i) = d_j, \tag{7}$$

$$\sum_j \pi_{ij} = (u_j\sum_i M_{ij})v_i = s_i. \tag{8}$$

当（7）（8）同时满足时，一个uv的可能解可以由一以下的迭代产生：

$$u_j^{t+1} = \frac{d_j}{\sum_i M_{ij}v_i^t}, \quad v_i^{t+1} = \frac{s_i}{\sum_j M_{ij}u_j^{t+1}}. \tag{9}$$

最终的近似最优解为：

$$\pi^{*} = diag(v)Mdiag(u) \tag{10}$$

EMD距离应用于feature map上时，两个特征图分别作为资源与目的地，那么损失可以定义为：

$$c_{ij} = 1 - \frac{\mathbf{x}_i^T\mathbf{y}_j}{\|\mathbf{x}_i\|\|\mathbf{y}_j\|}, \tag{11}$$

nodes with similar representations tend to generate

在计算得到最优的转换之后，可以得到两个图像特征图表示之间的相似度：

$$s(\mathbf{X},\mathbf{Y}) = \sum_{i=1}^{HW}\sum_{j=1}^{HW}(1 - c_{ij})\pi_{ij}^{*} \tag{12}$$

实验结果

Detector	Pre-train Method	Pre-train Data	Pre-train Label	AP	AP_{50}	AP_{75}	AP_s	AP_m	AP_l
FPN	Classification	ImageNet	✓	39.1	60.0	42.2	24.1	42.7	50.4
	BYOL	ImageNet		39.9	60.2	43.2	23.3	43.2	52.8
	Self-EMD	ImageNet		40.0	60.4	44.0	23.5	43.8	52.2
	BYOL	COCO		38.8	58.5	42.2	23.3	41.4	49.5
	Self-EMD	COCO		**39.8**	60.0	43.4	24.2	42.7	50.6
	BYOL	COCO+		39.3	59.0	42.8	23.5	42.1	50.5
	Self-EMD	COCO+		**40.4**	61.1	43.7	24.4	43.3	51.3
Mask+C4	Classification	ImageNet	✓	38.2	58.2	41.2	21.6	42.7	52.1
	BYOL	COCO		37.9	57.5	40.9	21.6	42.6	51.2
	Self-EMD	COCO		**38.5**	58.3	41.6	21.5	43.3	51.9
Mask+FPN	Classification	ImageNet	✓	38.9	59.6	42.7	23.7	42.6	52.0
	BYOL	COCO		38.5	58.9	41.7	22.8	41.9	49.3
	Self-EMD	COCO		**39.3**	60.1	42.8	24.4	42.7	49.9
RetinaNet	Classification	ImageNet	✓	37.1	56.1	39.7	22.9	41.1	48.4
	BYOL	COCO		36.2	54.8	38.8	21.3	40.2	45.4
	Self-EMD	COCO		**37.4**	56.5	39.7	23.2	41.3	46.6

Table 1. Main detection performance of several typical detectors in terms of AP (%) on COCO *val*. 'COCO+' denotes the COCO *train* 2017 set plus the COCO *unlabel* set. For fair comparison, BYOL [9] and *Self*-EMD are both pre-trained for 300 epochs on ImageNet and 800 epochs on COCO.

更多资料

- 微信公众号： 小哲AI

- GitHub地址： https://github.com/lxztju/leetcode-algorithm
- csdn博客： https://blog.csdn.net/lxztju
- 知乎专栏： https://www.zhihu.com/column/c_1101089619118026752
- AI研习社专栏：https://www.yanxishe.com/column/109