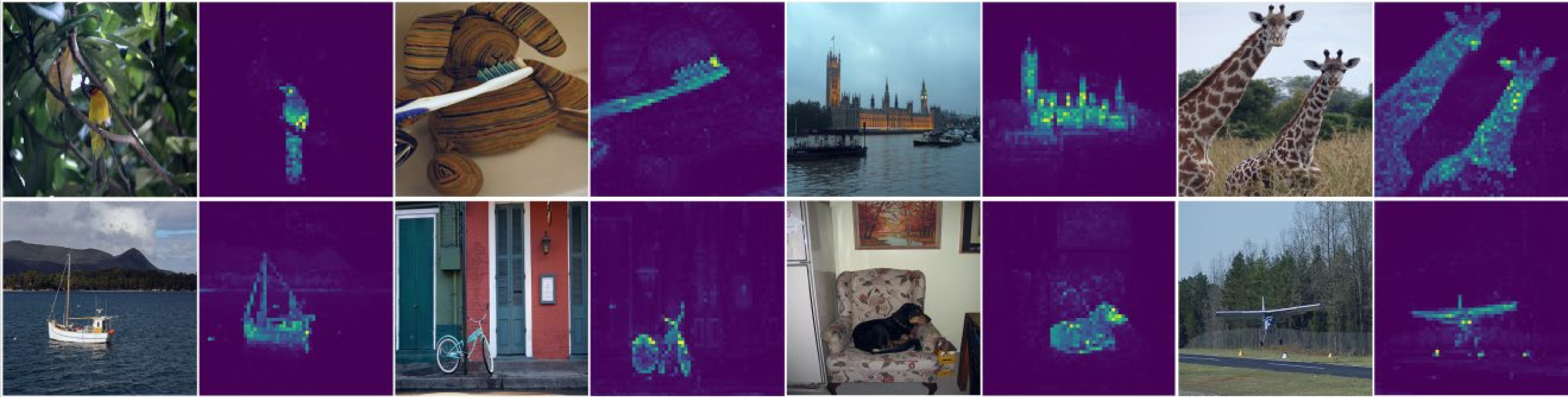
Emerging Properties in Self-Supervised Vision Transformers

https://arxiv.org/pdf/2104.14294.pdf

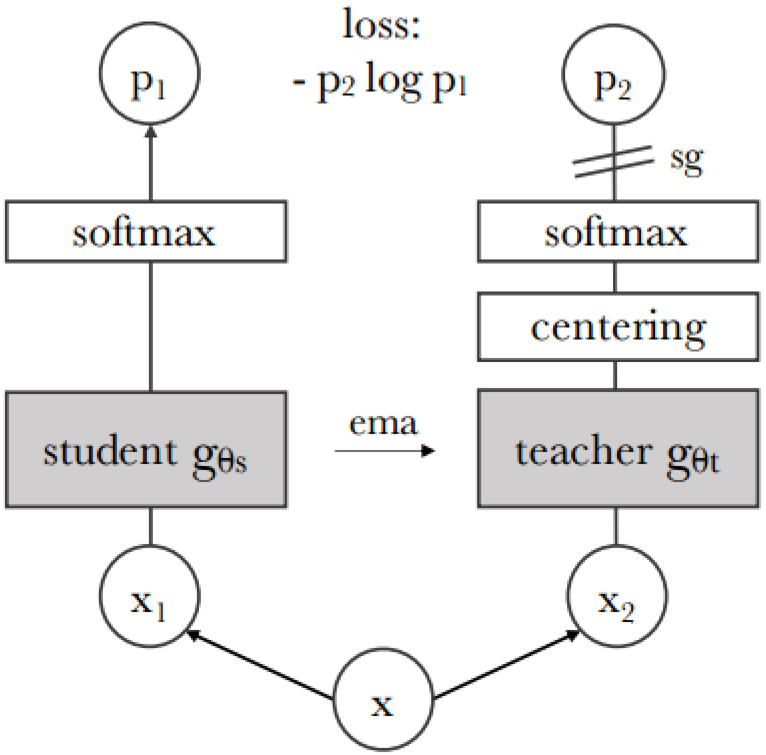
https://github.com/facebookresearch/dino

DINO



KEY WORDS:
1. No labels, No supervision
2. self-supervised *self-**di**stillation*
3. Achieving 80.1% top-1 on ImageNet

DINO



Key points:

1. 数据no label
2. 学生模型教师模型输入的数据不同
3. 教师模型不是预训练模型
4. 教师模型不通过梯度更新而是通过ema方式更新
5. 教师模型侧多出centering

结构

Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

1. 数据增强方式：随机裁剪出大图与小图、水平翻转、颜色变化、高斯模糊
2. 学生模型输入全部图像，教师模型输入大图，鼓励local-to-global
3.

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K}\exp(g_{\theta_s}(x)^{(k)}/\tau_s)},$$

$$\min_{\theta_s} \sum_{x\in\{x_1^g, x_2^g\}} \sum_{\substack{x'\in V \\ x'\neq x}} H(P_t(x), P_s(x')).$$

$H(a; b) = -a \log b$.

4. 学生模型参数更新
5. 教师模型参数通过学生模型ema更新

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$$

6. 更新centering

Avoiding collapse

Our framework can be stabilized with multiple normalizations, it can also work with only a centering and sharpening of the momentum teacher outputs to avoid model collapse. ***centering prevents one dimension to dominate but encourages collapse to the uniform distribution, while the sharpening has the opposite effect***.

$$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i),$$

m	0	0.9	0.99	0.999
k-NN top-1	69.1	69.7	69.4	0.1

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K}\exp(g_{\theta_s}(x)^{(k)}/\tau_s)},$$

τ_t	0	0.02	0.04	0.06	0.08	$0.04 \rightarrow 0.07$
k-NN top-1	43.9	66.7	69.6	68.7	0.1	69.7

消融实验

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**

Comparison across architectures

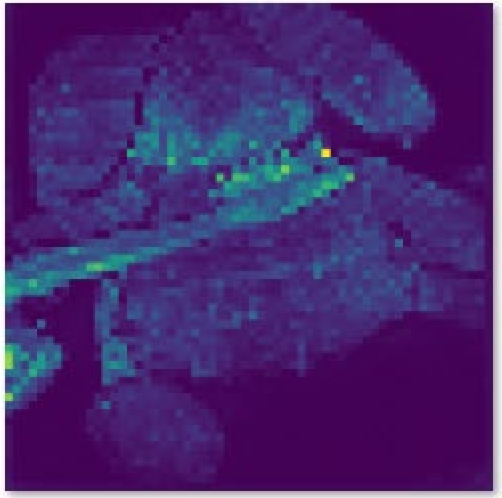
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [30]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

1. 相同的主干RN50下，DINO优于其他非监督学习模型
2. 不同主干下，vit更胜一筹
3. 相同主干vit下，patch size越小精度越高

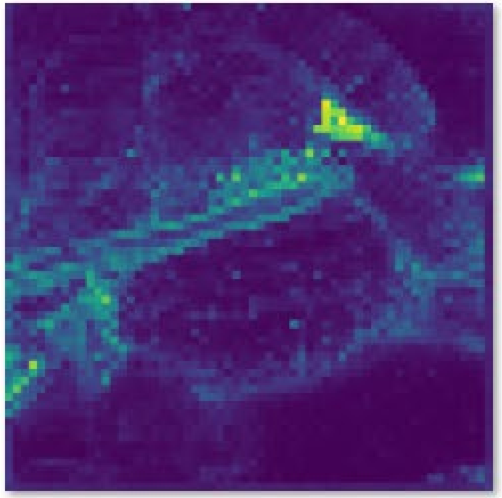
测试

	Random	Supervised	DINO
ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7

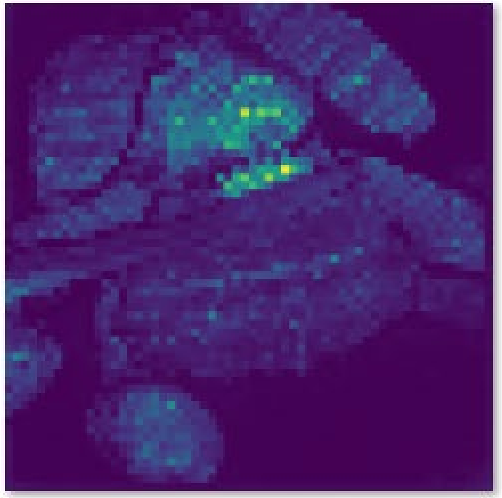
测试



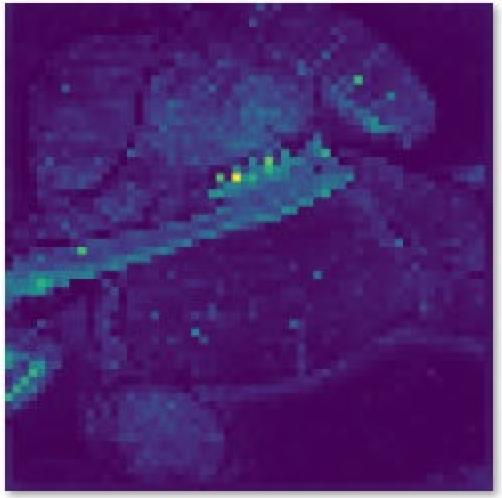
attn-head0.png



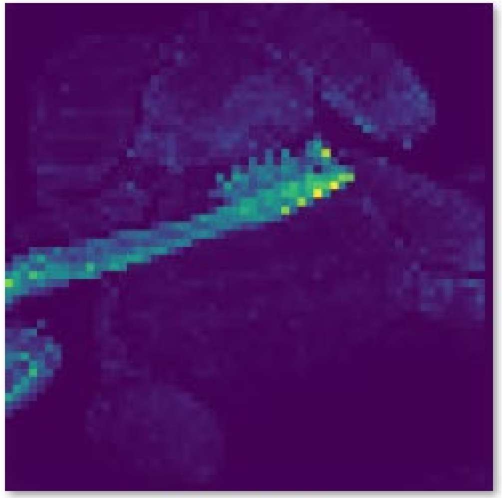
attn-head1.png



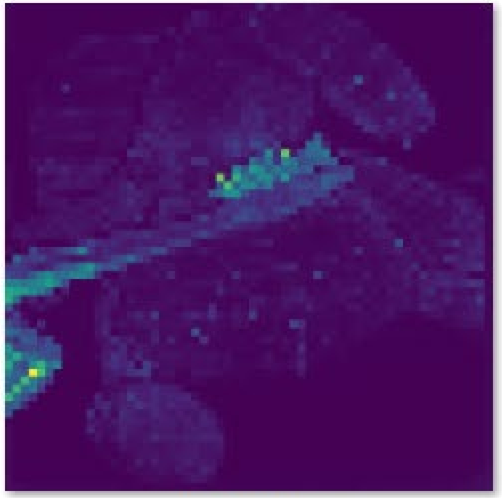
attn-head2.png



attn-head3.png

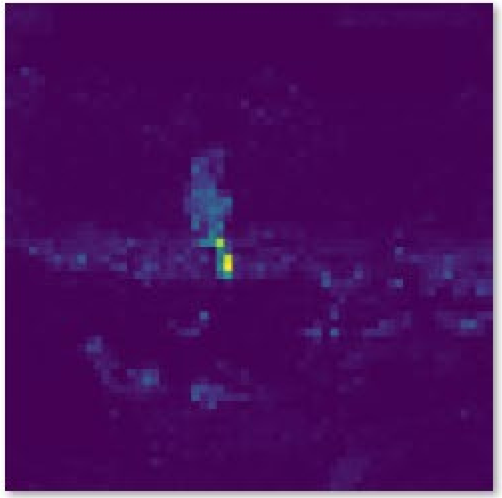


attn-head4.png

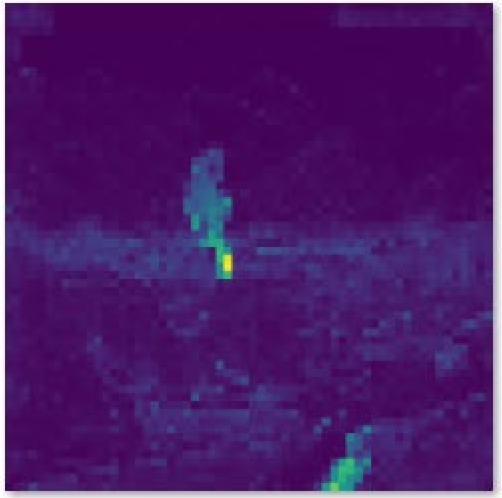


attn-head5.png

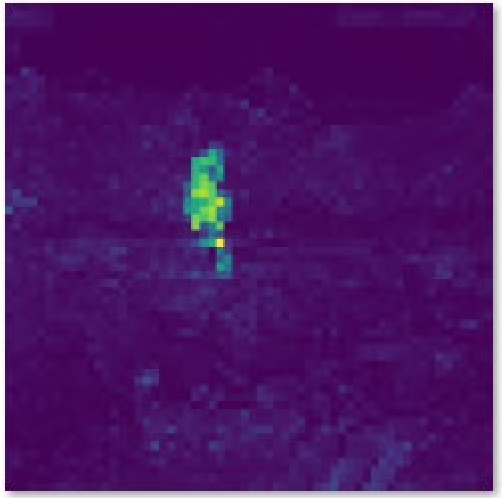
测试



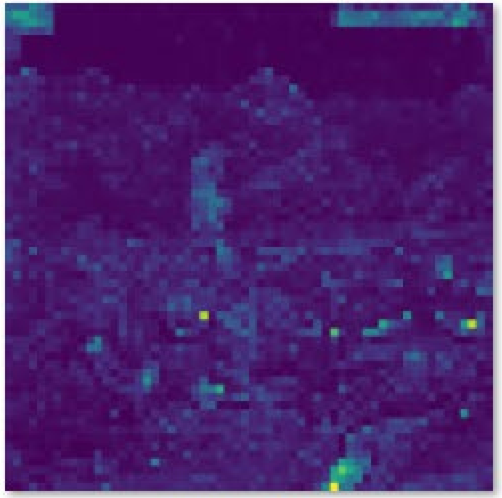
attn-head0.png



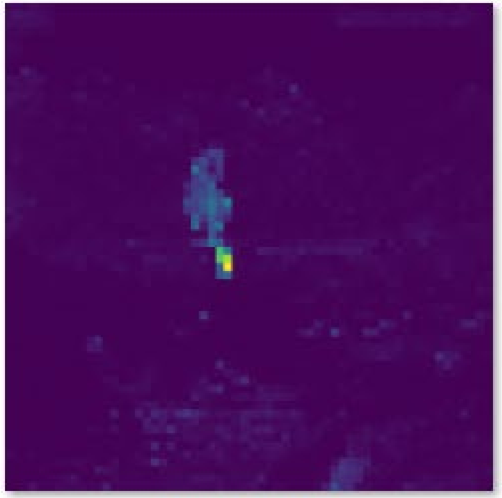
attn-head1.png



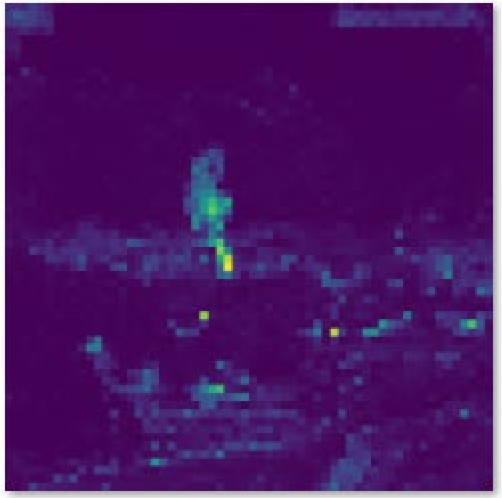
attn-head2.png



attn-head3.png



attn-head4.png



attn-head5.png