Note: Emerging Properties in Self-Supervised Vision Transformers

Sinkoo

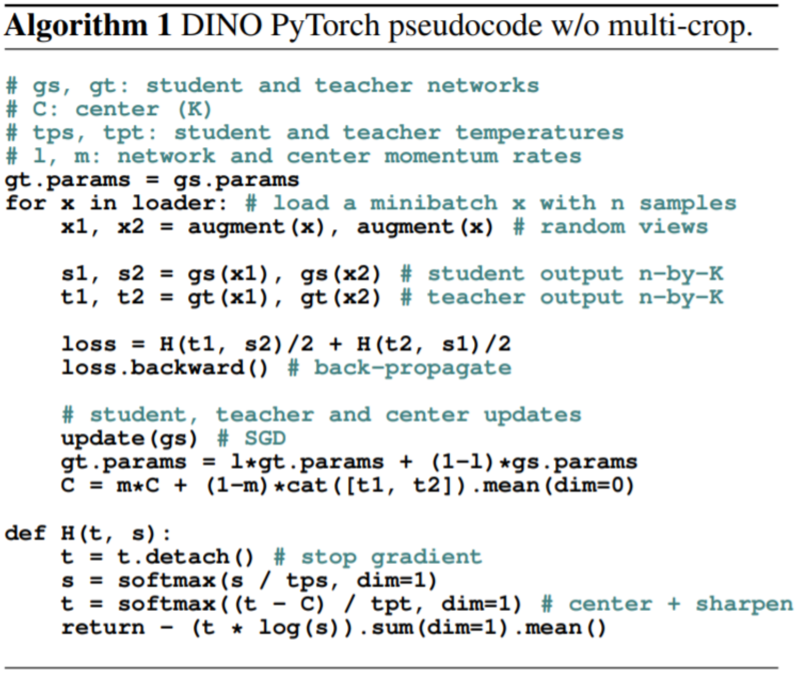
April 3, 2022

1 Abstract

In this paper, Author implemented self-supervision to ViT and works well, and they made the following observations: First, self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. Second, these features are also excellent k-NN classifiers.

They also found momentum encoder, multi-crop augmentation, and using smaller patches with ViTs to improve the quality of the resulting features important. And they developed a simple self-supervised method DINO. And get SOTA performance.

2 Method



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()
```

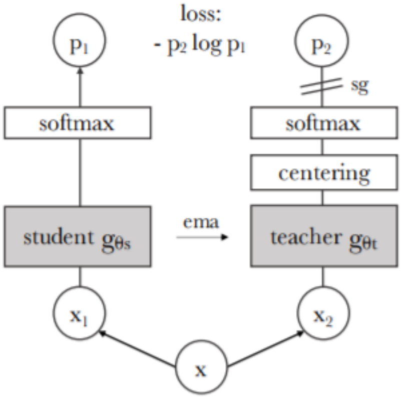


Figure 2: **Self-distillation with no labels.** We illustrate DINO in the case of one single pair of views (x_1, x_2) for simplicity. The model passes two different random transformations of an input image to the student and teacher networks. Both networks have the same architecture but different parameters. The output of the teacher network is centered with a mean computed over the batch. Each networks outputs a K dimensional feature that is normalized with a temperature softmax over the feature dimension. Their similarity is then measured with a cross-entropy loss. We apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters.

Knowledge Distillation For a fixed Net-T, train a Net-S to match the output of Net-T, parameterized by θ s and θ t respectively. Both Network output a soft distribution: $P_s(x)^{(i)} = \frac{exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} exp(g_{\theta_s}(x)^{(k)}/\tau_s)}$ τ is temperature that controls the sharpness of the output distribution. Minimize the cross-entropy.
In this paper, firstly constructing different distorted views or crops of an image with multicrop strategy, and view set(V) contains 2 global views(cover large area of image x_1^g, x_2^g) and several smaller local views(covering small area of image). All the corps are passed to Net-S and only global corps are passed to Net-T, therefore encouraging "local to global" correspondence(?By using information of local corps (Net-S) to predict global approximate output of Net-T which uses pure global corps?). then minimize the loss:

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

Through that the problem is adapted to self-supervised learning.

Teacher network The teacher network is based on the student network of precious epoch. Freezing the teacher network over an epoch works well. Using an exponential moving average (EMA) on the student weights is particularly well suited. The update rule is $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$, with λ following a cosine schedule from 0.996 to 1 during training.

Network architecture Network is composed of a backbone f (ViT or ResNet) and a projection head(a three layers MLP) $h : g = h \circ g$. Student and Teacher Network share the same architecture. features used in downstream tasks are the output of f.And when applying to ViT, DINO does not require BN like naive ViT.(BN-free).

Avoiding Collapse Applying centering and sharpening and balance their effect is sufficient to avoiding collapse.
Centering:Interpreted as a bias add to teacher $g_t(x) \leftarrow g_t(x) + c$ and upgrade through EMA :$c \leftarrow mc + (1-m)\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i)$ B is the batch size.
Sharpening: Use low value as τ_s

3 Ablation Study

1. Decreased patch size brings better performance at the expense of throughput.
2. Teacher Network: Copy of the student(precious iteration) for the teacher: does not converge. A momentum encoder works great.

4 Summary

This paper provided an efficient pretraining method with KD, pushed the limit of ViT. And it also showed the potential of self-supervised Learning in ViT pretraining. I think features obtained from pretraining methods like self-supervised pretraining could greatly increase efficiency of most of models in this field as many useless features will no longer hinder model's performance.

5 Other Knowledge