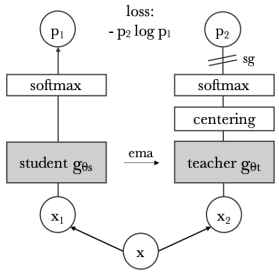
DINO v1

A Preprint

Authors take motivation from NLP where similar transformer architectures succeed by using self supervised techniques. It is mentioned that

- Self-supervised ViT features explicitly contain the scene layout and, in particular, object boundaries
- Self-supervised ViT features perform particularly well with a basic nearest neighbors classifier (k-NN) without any finetuning, linear classifier nor data augmentation, achieving 78.3% top-1 accuracy on ImageNet

The resulting framework, DINO, simplifies self-supervised training by directly predicting the output of a teacher network—built with a momentum encoder—by using a standard cross-entropy loss. Also, method can work with only a centering and sharpening of the teacher output to avoid collapse



Authors mention that recent work in this topic do not necessary need discriminating between images, such methods as BYOL where features are trained by matching them to representations obtained with a momentum encoder where momentum encoder as additional network which is used with sliding mean approach. The idea is that momentum encoder produces more stable weights but yet it takes long time to process so this particular encoder is actually used as teacher. The main factor is that teacher is trained the same time the student is trained or so called codistillation. For softmax which is used for logits authors use modification called temperature which is basically dividing by parameter t which takes repsonsibily for logit distrubition sharpness as if we divide by big t: distribution smoothens but if value is small we get sharp dist. Given a fixed teacher we learn n to match these distributions by minimizing the cross-entropy loss w.r.t. the parameters of the student network θ_s. Authors generate multiple views and use multi crop strategy (global views and local views. All crops are used for student and only-global ones for teacher). Authors follow the standard setting for multi-crop by using 2 global views at resolution 2242 covering a large (for example greater than 50%) area of the original image, and several local views of resolution 962 covering only small areas (for example less than 50%) of the original image. We refer to this setting as the basic parametrization of DINO, unless mentioned otherwise. It the paper mentioned that different teacher update techniques is studied, however only one works surprisingly well. Updating teacher over an epoch works way better then over iteraion or freezing. Also, EMA or exponentially moving average works well: $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_t$.

Neural network architecture is basically ViT backbone or ResNet and a projection head. The projection head consists of a 3-layer multi-layer perceptron (MLP) with hidden dimension 2048 followed by l_2 normalization and also fully connected layer with K dims.