DINO: distillation with no labels

① DINO : Distillation with No Labels.

Approach: SSL with KD:

Student Network g_{θ_s} ; teacher Net g_{θ_t} → K-dimentional Output.

Softmax conversion of student Output.

$$P_s(x)^i = \frac{\exp\{g_{\theta_s}(x)^i / \tau_s\}}{\sum_{i'=1}^{K} \exp\{g_{\theta_s}(x)^{i'} / \tau_s\}} \quad \text{// student}$$

$$P_t(x)^i = \frac{\exp\{g_{\theta_t}(x)^i / \tau_t\}}{\sum_{j=1}^{K} \exp\{g_{\theta_t}(x)^j / \tau_t\}} \quad \text{// teacher.}$$

Optimization goal: $\min_{\theta_s} H(P_t(x), P_s(x))$ // $H(a,b) = -a \log b$

Actually: Local-to-Global view

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

V → Set of images with different view of x.

$\{ x_1^g, x_2^g, x_1^L, \ldots x_n^L \}$

x_1^g, x_2^g ↓ global views { passed through teacher only }

$x_1^L \ldots x_n^L$ ↓ Local views { smaller dimentions } { student network }

Teacher Network Update:

$$\theta_t \leftarrow \lambda \theta_{t-1} + (1-\lambda)\theta_s \quad \text{// Polyak - Ruppert Avg}$$

Interesting Update.

Avoid collapse:

centering: $g_t(x) \leftarrow g_t(x) + c$ ← centering

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

Exp. Avg