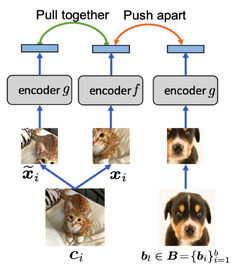
A Theory-Driven Self-Labeling Refinement Method for Contrastive Representation Learning

Pan Zhou* , Caiming Xiong* , Xiao-Tong Yuan† , Steven HOI*
* Salesforce Research † Nanjing University of Information Science & Technology
panzhou3@gmail.com {cxiong, shoi}@salesforce.com xtyuan@nuist.edu.cn

homepage

Problem Setup

MoCo pulls together crops of same image, pushs away crops of different images



Given a minibatch samples $\{\boldsymbol{c}_i\}_{i=1}^s$, it augments each $\boldsymbol{c}_i$ into two views $(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i)$ and optimizes:

$$\mathcal{L}_{\rm n}(\boldsymbol{w}) = -\frac{1}{s}\sum_{i=1}^{s} \log\Big(\frac{\sigma(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i)}{\sigma(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i) + \sum_{l=1}^{b}\sigma(\boldsymbol{x}_i, \boldsymbol{b}_l)}\Big),$$

query · positive key · negative key

- $\sigma(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i) = \exp\big(-\frac{\langle f(\boldsymbol{x}_i), g(\widetilde{\boldsymbol{x}}_i)\rangle}{\tau\|f(\boldsymbol{x}_i)\|_2 \cdot \|g(\widetilde{\boldsymbol{x}}_i)\|_2}\big)$: cosine similarity function
- $f_{\boldsymbol{w}}$: online network
- g_ξ: target network
- $\boldsymbol{B} = \{\boldsymbol{b}_l\}_{l=1}^b$: negative key buffer (previous minibatch crops)

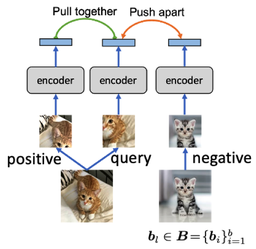
One-hot label assignment: query $\boldsymbol{x}_i$ has only one positive $\widetilde{\boldsymbol{x}}_i$ among $\widetilde{\boldsymbol{x}}_i \cup \boldsymbol{B}$

$$\mathcal{L}_{\rm n}(\boldsymbol{w}) = -\frac{1}{s}\sum_{i=1}^{s} \log\Big(\frac{\sigma(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i)}{\sigma(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i) + \sum_{l=1}^{b}\sigma(\boldsymbol{x}_i, \boldsymbol{b}_l)}\Big),$$

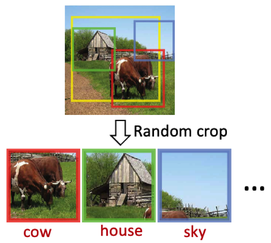
query · positive key · negative key

Issues of hot label assignment: imprecise & uninformative:

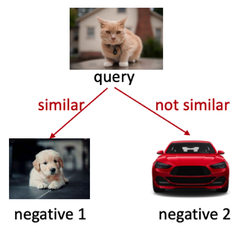
(1) some negatives & query from same semantic class



(2) augmentations gives different semantic crops



(3) negatives have different similarity to query



Result: one-hot label cannot guarantee semantically similar samples to close

Theoretical Motivation

Support that the pair $(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i)$ in the training dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i)\}_{i=1}^n$ sampled from an unknown distribution $\mathcal{S}$ denotes the positive pair in MoCo.
Assume the query $\boldsymbol{x}_i$ has ground truth soft label $\boldsymbol{y}_i^* \in \mathbb{R}^{b+1}$ over the key set $\boldsymbol{B}_i = \{\widetilde{\boldsymbol{x}}_i \cup \boldsymbol{B}\}$ where $\boldsymbol{y}_{it}^*$ measures the semantic similarity between $\boldsymbol{x}_i$ and the t-th key $\boldsymbol{b}_t'$ in buffer $\boldsymbol{B}_i$

Theorem 1 (upper bound of generalization error, informal).
Under proper assumptions, for MoCo, with probability $1-\nu$, the generalization error on instance discrimination task can be upper bounded as :

one-hot label · true soft label

$$\text{generalization error} \le \mathcal{O}\big(\mathbb{E}_{\mathcal{D}\sim\mathcal{S}}\,[\|\boldsymbol{y} - \boldsymbol{y}^*\|_2]\big) + \mathcal{O}\Big(\sqrt{\frac{V_{\mathcal{D}}\ln(|\mathcal{F}|/\nu)}{n}} + \frac{\ln(|\mathcal{F}|/\nu)}{n}\Big),$$

training & test error gap

where $V_{\mathcal{D}}$ is the variance of $f_{\boldsymbol{w}}$ on data $\mathcal{D}$, $\mathcal{F}$ is the covering number of encoder $f_{\boldsymbol{w}}$

Remark: **the more accurate of the label y, the better the generalization**

Theoretical Motivation

Theorem 2 (lower bound of generalization error, informal).
Under proper assumptions, for MoCo, there exists a contrastive learning problem such that the generalization error on instance discrimination task is lower bounded as :

one-hot label · true soft label

$$\text{generalization error} \ge \mathcal{O}\big(\mathbb{E}_{\mathcal{D}\sim\mathcal{S}}\,[\|\boldsymbol{y} - \boldsymbol{y}^*\|_2]\big).$$

Remark: **lower & upper bounds show generalization error** $\sim \mathbb{E}_{\mathcal{D}\sim\mathcal{S}}\,[\|\boldsymbol{y} - \boldsymbol{y}^*\|_2]$
the more accurate of the label y, the better the generalization

Solution: Self-Labeling Refinement

Self-Labeling Refinement has two components:

- **self-labeling refinery**: soft label replaces one-hot label to directly improve label accuracy
- **momentum mixup**: increase similarity of positive pair to indirectly improve label accuracy

MoCo Reformulation: for query $\boldsymbol{x}_i$ in minibatch $\{(\boldsymbol{x}_i, \widetilde{\boldsymbol{x}}_i)\}_{i=1}^s$, we maximize its similarity to its positive $\boldsymbol{x}_i$ in key set $\bar{\boldsymbol{B}} = \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s \cup \{\boldsymbol{b}_i\}_{i=1}^b$ & push it away remaining samples:

$$\mathcal{L}_{\rm c}(\boldsymbol{w}, \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}) = -\frac{1}{s}\sum_{i=1}^{s}\sum_{k=1}^{s+b} \boldsymbol{y}_{ik} \log\Big(\frac{\sigma(\boldsymbol{x}_i, \bar{\boldsymbol{b}}_k)}{\sum_{l=1}^{s+b}\sigma(\boldsymbol{x}_i, \bar{\boldsymbol{b}}_l)}\Big),$$

where $\bar{\boldsymbol{b}}_k$ is the k-th sample in $\bar{\boldsymbol{B}}$, the i-th entry $\boldsymbol{y}_{ij}$ of one-hot label $\boldsymbol{y}_i$ of query $\boldsymbol{x}_i$ is 1.

Target of Reformulation: labels of different samples are defined on a shared dictionary, and thus can be linearly combined.

Self-labeling refinery iteratively uses network to improve labels during training

- Step1. for query $\boldsymbol{x}_i$, we use its positive $\widetilde{\boldsymbol{x}}_i$ to estimate semantic similarity between $\boldsymbol{x}_i$ and instances in $\bar{\boldsymbol{B}} = \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s \cup \{\boldsymbol{b}_i\}_{i=1}^b$, since $\boldsymbol{x}_i$ and $\widetilde{\boldsymbol{x}}_i$ come from the same image:

$$\boldsymbol{p}_{ik}^t = \sigma^{1/\tau'}(\widetilde{\boldsymbol{x}}_i, \bar{\boldsymbol{b}}_k) \big/ \sum_{l=1}^{s+b}\sigma^{1/\tau'}(\widetilde{\boldsymbol{x}}_i, \bar{\boldsymbol{b}}_l),$$

- Step2. as $\widetilde{\boldsymbol{x}}_i$ is highly similar to itself in $\bar{\boldsymbol{B}}$, $\boldsymbol{p}_{ii}^t$ is much larger than others, conceals similarity of other semantically similar instances in $\bar{\boldsymbol{B}}$. So we remove $\widetilde{\boldsymbol{x}}_i$ from $\bar{\boldsymbol{B}}$

$$\boldsymbol{q}_{ik}^t = \sigma^{1/\tau'}(\widetilde{\boldsymbol{x}}_i, \bar{\boldsymbol{b}}_k) \big/ \sum_{l=1, l\ne i}^{s+b}\sigma^{1/\tau'}(\widetilde{\boldsymbol{x}}_i, \bar{\boldsymbol{b}}_l), \quad \boldsymbol{q}_{ii}^t = 0.$$

- Step3. Linear combination

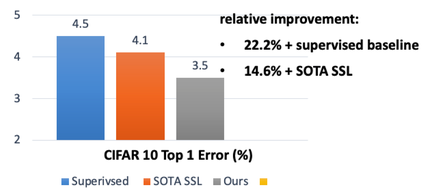
$$\bar{\boldsymbol{y}}_i^t = (1-\alpha_t-\beta_t)\boldsymbol{y}_i + \alpha_t \boldsymbol{p}_i^t + \beta_t \boldsymbol{q}_i^t,$$

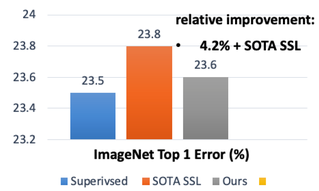
Momentum mixup constructs virtual instance as follows:

$$\boldsymbol{x}_i' = \theta\boldsymbol{x}_i + (1-\theta)\widetilde{\boldsymbol{x}}_k, \quad \boldsymbol{y}_i' = \theta\bar{\boldsymbol{y}}_i + (1-\theta)\bar{\boldsymbol{y}}_k, \quad \theta \sim \text{Beta distribution}$$

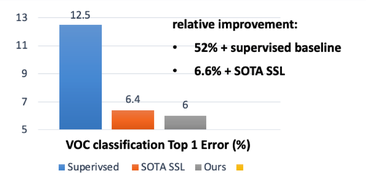
where $\widetilde{\boldsymbol{x}}_k \overset{random}{\sim} \{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s$, $\bar{\boldsymbol{y}}_i$ is refined label by self-labeling refinery

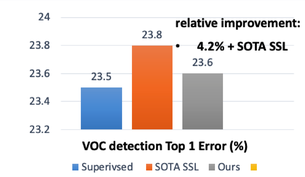
Benefits: the component $\widetilde{\boldsymbol{x}}_k$ in $\boldsymbol{x}_i'$ increases similarity between query $\boldsymbol{x}_i'$ and positive key $\boldsymbol{x}_i'$, which improves the label accuracy





Experiments





Theoretical Analysis on Self-Labeling Refinary

Label-corrupted dataset

Let $\{(\boldsymbol{x}_i, \boldsymbol{y}_i^*)\}_{i=1}^n$ resp. denote the pairs of crops and ground-truth semantic label

- crop $\boldsymbol{x}_i$ generated from vanilla sample $\boldsymbol{c}_t$ obeys $\|\boldsymbol{x}_i - \boldsymbol{c}_t\|_2 \le \varepsilon$
- ground-truth semantic label $\boldsymbol{y}_i^* \in \{\gamma_t\}_{t=1}^K$ of $\boldsymbol{x}_i$ is decided by its corresponding $\boldsymbol{c}_t$
- the classes are separated: $|\gamma_i - \gamma_k| \ge \delta, \quad \|\boldsymbol{c}_i - \boldsymbol{c}_k\|_2 \ge 2\varepsilon, \quad (\forall i \ne k)$,
- for each sample $\boldsymbol{c}_i$, at most ρn_i augmentations are assigned to wrong labels, where n_i denotes the crop sample number of $\boldsymbol{c}_i$

Assume online/target networks f/g: $\boldsymbol{x} \in \mathbb{R}^d \mapsto f(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{v}^\top\phi(\boldsymbol{W}\boldsymbol{x})$
training loss: $\mathcal{L}_t(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^n(\bar{\boldsymbol{y}}_i^t - f(\boldsymbol{W}, \boldsymbol{x}_i))^2 = \frac{1}{2}\|\bar{\boldsymbol{y}}^t - f(\boldsymbol{W}, \boldsymbol{X})\|_2^2$
gradient descent algorithm: $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta\nabla\mathcal{L}_t(\boldsymbol{W}_t)$

Theorem 3 (exact label recovery on training data, informal).
Under proper assumptions, after t iterations, for data $\{\boldsymbol{x}_i\}_{i=1}^n$, we have

label corrupted ratio

$$\frac{1}{\sqrt{n}}\|\boldsymbol{y}^t - \boldsymbol{y}^*\|_2 \le \frac{1-\alpha_t}{\sqrt{n}}\|\boldsymbol{y}^0 - \boldsymbol{y}^*\|_2 + \alpha_t(6\rho + \zeta),$$

estimated label · true soft label

where $\boldsymbol{y}^t = [\boldsymbol{y}_1^t, \cdots, \boldsymbol{y}_n^t]$, $\boldsymbol{y}^* = [\boldsymbol{y}_1^*, \cdots, \boldsymbol{y}_n^*]$, $\zeta = c\varepsilon K^2\sqrt{\log K}$ is related to network

Moreover, if the following two assumption hold,

- $\rho \le \frac{\delta}{24}$: label corrupted ratio is small
- $(1-\alpha_0)|\boldsymbol{y}_i^0 - \boldsymbol{y}_i^*| + \frac{1}{3}\alpha_0\delta < \frac{1}{2}\delta$: label noise magnitude is small (δ is class label separation)

the estimated label $\boldsymbol{y}_i^t$ predicts true label $\boldsymbol{y}_i^*$ of any crop $\boldsymbol{x}_i$

Exact label recovery: $\gamma_{k^*} = \boldsymbol{y}_i^* \quad \text{with} \quad k^* = \text{argmin}_{1\le k\le K}|\boldsymbol{y}_i^t - \gamma_k|$.
estimated label · true label

Remark: along training, self-label refinery recovers corrupted training labels

Theorem 4 (exact prediction of network trained by self-label refinery, informal)
Under proper assumptions, by using the refined label $\boldsymbol{y}^t$ to train network, the error of network prediction on **training data** $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ is upper bounded as

$$\frac{1}{\sqrt{n}}\|f(\boldsymbol{W}_t, \boldsymbol{X}) - \boldsymbol{y}^*\|_2 \le 6\rho + \frac{c\zeta}{K\Gamma^2},$$

predicted label · true label

Moreover, if the following two assumption hold,

- $\rho \le \frac{\delta}{24}$: label corrupted ratio is small
- $(1-\alpha_0)|\boldsymbol{y}_i^0 - \boldsymbol{y}_i^*| + \frac{1}{3}\alpha_0\delta < \frac{1}{2}\delta$: label noise magnitude is small (δ is class label separation)

for any vanilla sample $\boldsymbol{c}_k$, network $f(\boldsymbol{W}_t, \cdot)$ predicts true label $\boldsymbol{y}_i^*$ of any **test augmentation** $\boldsymbol{x}$ that obeys $\|\boldsymbol{x} - \boldsymbol{c}_k\|_2 \le \varepsilon$:

Exact label prediction: $\gamma_{k^*} = \gamma_k \quad \text{with} \quad k^* = \text{argmin}_{1\le i\le K}|f(\boldsymbol{W}_t, \boldsymbol{x}) - \gamma_i|$.
estimated label · true label

Remark: network trained by self-label refinery predicts label well