Not All Semantics are Created Equal: Contrastive Self-supervised Learning with Automatic Temperature Individualization

Zi-Hao, Qiu; Quanqi, Hu; Zhuoning, Yuan; Denny, Zhou; Lijun, Zhang; Tianbao, Yang

Texas A&M, Nanjing U, U of Iowa, Google Research

Jan, 2025

Presented by Scott Sun from Duke B&B

Introduction

The authors developed a thoroughly-analyzed methodological approach to optimize InfoNCE-like contrastive loss with individualized temperature. They proposed the idea that "not all semantics are created equal": samples with frequent semantics should be assigned with a **large** τ, while samples with infrequent semantics should be assigned with a **small** τ.

Goal: allow individualized temperature to address representation learning problems that arise in long-tail data distribution, where semantics with low frequencies can be extremely diverse. Therefore, address the hard negatives in a long-tail data distribution.

- **semantics harmonizing effect:** for global τ achieve the a sweet spot is hard; large τ is better at capturing local semantic structures in more frequent classes (but can be excessively forgiving for semantic difference); small τ is better at capturing more discriminative and separable features (but can result in fragmented semantic stuctures)
- **iSogCLR**: a stochastic optimization algorithm inspired by DRO for robust contrastive learning with temperature individualization

Background: Current Methods

Contrastive Learning (CL) methods are nowadays very popular for Self-supervised Learning (SSL). Various great works have been made for both unimodal and bimodal tasks, including SimCLR, MoCo, CLIP, etc. Optimizing τ in CL can be challenging. Currently, there are 2 different ways to incorporate learnable temperature:

1. universal learnable hyperparameter (i.e., CLIP)
2. fixed values/input-dependent functions (i.e., TaU)

Both are not justified in theory and have obvious drawbacks. CLIP is based on a universal temperature, so it still ignores the imbalanced semantics problem. TaU, by taking input-dependent temperatures as uncertainty measures, is effective in OOD detection but sacrifice downstream performance.

Background: Notations

Notation	Description
$\mathcal{D} = \{x_1, ..., x_n\}$	entire training set
$\mathcal{P}$	set of augmentation operator
$\mathcal{S}_i^- = \{\mathcal{A}(x) : \forall \mathcal{A} \in \mathcal{P}, \forall x \in \mathcal{D} \backslash \{x_i\}\}$	negative set with anchor image x_i
$E(\cdot)$	modality encoder
$\Delta_n = \{\boldsymbol{p} : \sum_j p_j = 1, \forall j, p_j \geq 0\}$	simplex of dim n

Table: Notation and their descriptions.

Following the same manner, we have similar notations for bimodal tasks: $\mathcal{I}_i^-, \mathcal{T}_i^-, E_I(\cdot), E_T(\cdot)$

Background: Global Contrastive Loss (GCL)

Let's recall the NT-Xent loss's gradient w.r.t. latent embedding of x_i.

$$\nabla l_{\text{NT-Xent}} = \frac{1}{\tau}\left[\left(1 - \frac{\exp\left(E(\mathcal{A}(x_i))^\top E(\mathcal{A}'(x_i))/\tau\right)}{Z(x_i)}\right) \cdot E(\mathcal{A}'(x_i)) + \sum_{z \in S_i^- \cup \{\mathcal{A}'(x_i)\}} \frac{\exp\left(E(\mathcal{A}(x_i))^\top E(z)/\tau\right)}{Z(x_i)} \cdot E(z)\right] \tag{1}$$

where $Z(x_i) = \sum_z \exp\left(E(\mathcal{A}(x_i))^\top E(z)/\tau\right)$. The gradient can be illy scaled by extreme values of τ, especially when we considering individualized temperatures. A more robust contrastive loss called GCL is then defined by Yuan et al. (2022).

$$l_{\text{GCL}}(x_i) = -\tau \log \frac{E(\mathcal{A}(x_i))^\top E(\mathcal{A}'(x_i))/\tau}{\sum_{z \in S_i^-} E(\mathcal{A}(x_i))^\top E(z)/\tau} \tag{2}$$

which rescales the $l_{\text{NT-Xent}}$ by τ and remove positive sample from the denominator.

Background: Global Contrastive Loss (GCL) cont'd

Let $h_i(z) := E(\mathcal{A}(x_i))^\top E(z) - E(\mathcal{A}(x_i))^\top E(\mathcal{A}'(x_i))$ (which can be seen as measure for the hardness of sample z). Then,

$$l_{\text{GCL}}(x_i) = \tau \log \sum_{z \in S_i^-} \exp(h_i(z)/\tau) \tag{3}$$

A hard negative sample closely resembles positive samples in its latent representation, making it more challenging to distinguish from them. Therefore, a high $h_i(z)$ (i.e. 0) means its almost identical to the anchor sample in the latent space.

Why we want to use a global loss? Mini-batch-based algorithms (i.e. SimCLR) are vulnerable to the choice batch size. For ImageNet-like datasets with hundreds of classes, the batch size is suggested to be 8,192 in order to achieve a satisfied performance. A global loss can be more robust to the batch size. Still, we have to convert GCL to a stochastic algorithm that can be run over mini-batches.

Background: DRO

General DRO formulation is given by Levy et al., 2020:

$$\min_{\boldsymbol{w}} \max_{\boldsymbol{p} \in \mathcal{U}} \sum_{i=1}^{n} p_i l_i(\boldsymbol{w}) - \lambda D(\boldsymbol{p}, \mathbf{1}/n) \tag{4}$$

where $\mathcal{U} \subset \Delta_n$ is the uncertainty set of DRO.
In this setting, we optimize a weighted aggregated loss that is robust to distribution uncertainy.

Maximizing the objective over $\boldsymbol{p}$ leads to larger weights on samples with larger losses, which finds the worst-case loss (say, hard to contrast). DRO then minimizes the worst-case loss to make models achieve the robustness against potential distribution shifts.

Method: From DRO to Robust Global Contrastive Loss (RGCL)

Formulating under the previous DRO setup,

$$l_{\mathrm{RGCL}}(x_i) = \max_{\boldsymbol{p}\in\Delta_m} \sum_{z_j \in \mathcal{S}_i^-} p_j h_i(\boldsymbol{z}_j) - \tau_0 \mathrm{KL}(\boldsymbol{p}, \mathbf{1}/m) \quad \text{s.t.} \quad \mathrm{KL}(\boldsymbol{p}, \mathbf{1}/m) \leq \rho \tag{5}$$

where it can be shown that $\boldsymbol{p}_j^* \propto \exp\left(h_i(\boldsymbol{z}_j)/\tau\right)$ in our context.
Applying Lagrangian multiplier (λ) and Lagrangian duality theory, we can achieve the dual objective

$$\min_{\tau \geq \tau_0} \log \mathbb{E}_{\boldsymbol{z}\in\mathcal{S}_i^-} [\exp(h_i(z)/\tau)] + (\tau - \tau_0)\rho \tag{6}$$

where $\tau = \lambda + \tau_0$. Ultimately,

$$l^*_{\mathrm{RGCL}}(x_i) = F(\boldsymbol{w}, \boldsymbol{\tau}) = \frac{1}{n} \sum_{x_i \in \mathcal{D}} \left\{ \tau_i \log \mathbb{E}_{\boldsymbol{z}\in\mathcal{S}_i^-} \left[\exp\left(\frac{h_i(z)}{\tau} \right) \right] + \tau_i \rho \right\} \tag{7}$$

where τ_i is the individualized temperature in CL (implicitly, λ_i becomes individualized).

Method: RGCL is hardness-awaring

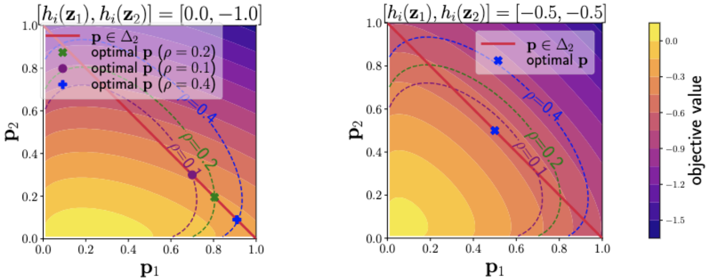


Figure 2. Contours of $\ell_{\mathrm{RGCL}}(\mathbf{x}_i)$ ($|\mathcal{S}_i^-| = 2$) in (6) for two $h_i = [h_i(\mathbf{z}_1), h_i(\mathbf{z}_2)]$ vectors: $[0.0, -1.0]$ and $[-0.5, -0.5]$. One can observe that $\ell_{\mathrm{RGCL}}(\mathbf{x}_i)$ is hardness-aware, harder sample ($h_i(\mathbf{z}_1)$ on the left) has larger weight ($\mathbf{p}_1 = 0.8$). Moreover, ρ affects the degree of hardness-awareness. Larger ρ means higher degree of hardness-awareness.

Method: RGCL is hardness-awaring

For samples with frequent semantics, $\boldsymbol{p}$ will tend to be non-uniform (see the left panel), the restriction will lead to a large λ and thus a large τ_i.

For samples with rare semantics, $\boldsymbol{p}$ will tend to be uniform (see the right panel), we do not need a large λ to obey the restriction. Hence, we have a small τ_i.

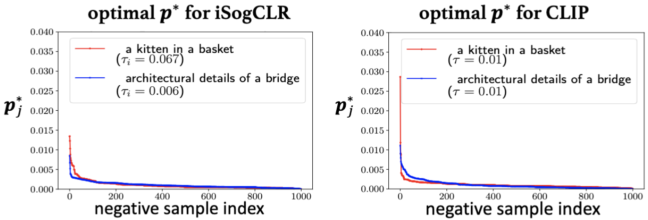


Figure 3. For the anchor images of cat and bridge, we select 1000 negative samples and solve (6) for the optimal $\mathbf{p}^*$ by using h_i values of iSogCLR with learned $\boldsymbol{\tau}_i$ and CLIP with learned τ.

Method: Stochastic Optimization iSogCLR algorithm

Although RGCL is developed for global optimization instead of mini-batch optimization, we have to adapt it to a mini-batch algorithm for practical implementation. We will not use mini-batch estimator for the loss function, since this will reduce the algorithm back to the mini-batch-based SimCLR format (requiring large batch size). Instead, we use moving average estimator.

Let $g_i(\boldsymbol{w}, \boldsymbol{\tau}; \mathcal{S}_i^-) = \exp(h_i(z)/\tau)$, then the moving average estimator of g_i is

$$s_i^{(t+1)} = (1 - \beta_0)s_i^{(t)} + \beta_0 g_i(\boldsymbol{w}^{(t)}, \tau_i^{(t)}; \mathcal{B}_i) \tag{8}$$

The new mini-batch gradients become

$$G(\tau_i^{(t)}) = \frac{1}{n}\left[\frac{\tau_i^{(t)}}{s_i^{(t)}}\nabla_{\tau_i} g_i(\boldsymbol{w}^{(t)}, \tau_i^{(t)}; \mathcal{B}_i) + \log(s_i^{(t)}) + \rho\right] \tag{9}$$

$$G(\boldsymbol{w}^{(t)}) = \frac{1}{B}\sum_{\mathbf{x}_i \in \mathcal{B}} \frac{\tau_i^{(t)}}{s_i^{(t)}}\nabla_{\boldsymbol{w}^{(t)}} g_i(\boldsymbol{w}^{(t)}, \tau_i^{(t)}; \mathcal{B}_i) \tag{10}$$

Experiment: Unimodal

Table 1. Linear evaluation results with 400 pretraining epochs on six unimodal image datasets. We report the average top-1 accuracies (%) and standard deviation over 3 runs with different random seeds. Full results are provided in Table 3 and 4 in Appendix C.3.

METHOD	CIFAR10	CIFAR100	IMAGENET100	CIFAR10-LT	CIFAR100-LT	INATURALIST
SIMCLR	88.74±0.18	62.34±0.09	79.96±0.20	77.09±0.13	49.33±0.12	91.52±0.17
BARLOW TWINS	87.39±0.14	62.28±0.13	79.16±0.13	75.94±0.08	48.39±0.14	91.89±0.21
FLATCLR	88.61±0.10	63.27±0.07	80.24±0.16	77.96±0.12	52.61±0.06	92.54±0.09
SPECTRAL CL	88.77±0.09	63.06±0.18	80.48±0.08	76.38±0.21	51.86±0.16	92.13±0.16
SOGCLR	88.93±0.11	63.14±0.12	80.54±0.14	77.70±0.07	52.35±0.08	92.60±0.08
VICREG	88.96±0.16	62.44±0.13	80.16±0.22	75.05±0.09	48.43±0.13	93.03±0.14
SIMCO	88.86±0.12	62.67±0.06	79.73±0.17	77.71±0.13	51.06±0.09	92.10±0.12
ISOGCLR	**89.24**±0.15	**63.82**±0.14	**81.14**±0.19	**78.37**±0.16	**53.06**±0.12	**93.08**±0.19

Experiment: Unimodal Temp Analysis

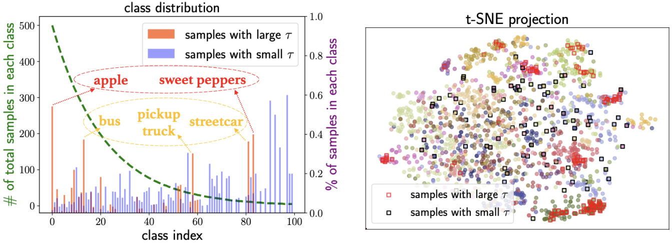


Figure 5. The class distributions and t-SNE projection for samples with large and small τ values in CIFAR100-LT. Left: The green dashed line and left axis denote the number of samples in each class, the red/blue bars and right axis denote the proportions of samples with large/small τ values in each class. Right: Each color represents a *superclass* in CIFAR100-LT.

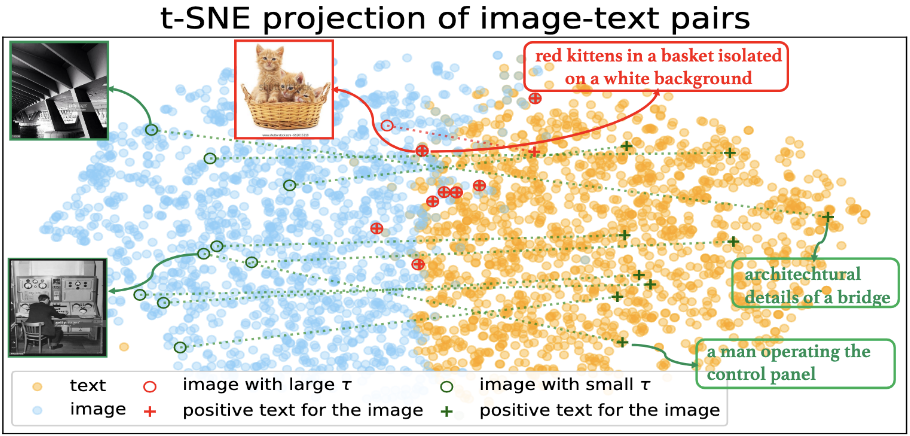
Clinical concerns: rare but severe disease might be misclassified as some commonly observed illness

Experiment: Bimodal

Table 2. Results on two bimodal downstream tasks. For image-text retrieval on Flickr30K and MSCOCO, we compute IR@1 and TR@1 for the Recall@1 on image-retrieval (IR) and text-retrieval (TR). For classification tasks, we compute top-1 accuracy (%). We report the average of scores and standard deviation over 3 runs with different random seeds. Full results are in Table 5, 6, and 7 in Appendix C.3.

Method	Flickr30K Retrieval		MSCOCO Retrieval		Zero-shot Classification top-1 Acc		
	IR@1	TR@1	IR@1	TR@1	CIFAR10	CIFAR100	ImageNet1K
CLIP	40.98±0.22	50.90±0.17	21.32±0.12	26.98±0.21	60.63±0.19	30.70±0.11	36.27±0.17
CyCLIP	42.46±0.13	51.70±0.23	21.58±0.19	26.18±0.24	57.19±0.20	33.11±0.14	36.75±0.21
SogCLR	43.32±0.18	57.18±0.20	22.43±0.13	30.08±0.22	**61.09**±0.24	33.26±0.12	37.46±0.19
iSogCLR	**44.36**±0.12	**60.20**±0.26	**23.27**±0.18	**32.72**±0.13	58.91±0.15	**33.81**±0.18	**40.72**±0.23

Experiment: Temp Analysis



Experiment: Ablation Study

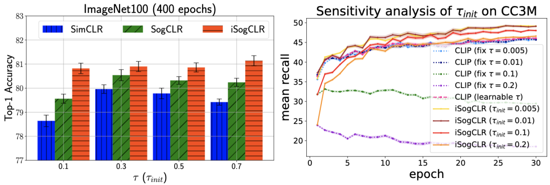


Figure 7. Effect of τ and τ_{init} on SimCLR/SogCLR and iSogCLR.

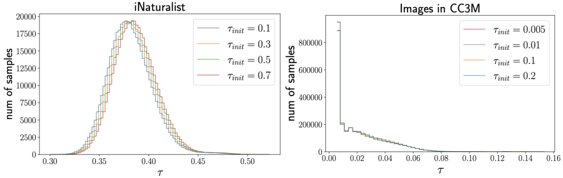


Figure 8. Final distributions of learned temperatures.

Recommendation

Is it worth reading? Yes.

- the math/framework is thorough and elegant; the experimental interpretation is straightforward
- the paper gives clear illustration of how the new RGCL is developed from DRO

Is it worth implementing? Yes.

- they have a public github repo which includes the code to reproduce all the results, but the documentation is minimal
- would love to see how this would improve my current project about semantic disentanglement