On the Surrogate Gap between Contrastive and Supervised Losses

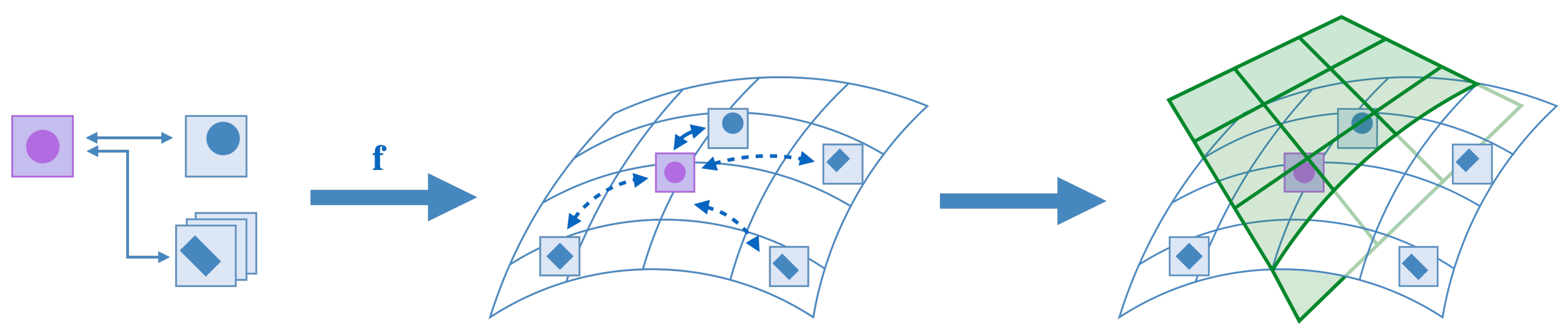
Han Bao*[1,2] Yoshihiro Nagano*[1,2] Kento Nozawa*[1,2]

[1]The University of Tokyo, Japan / [2]RIKEN AIP, Japan / *Equal contribution

arXiv: 2110.02501

Introduction

Contrastive Unsupervised Representation Learning (CURL)

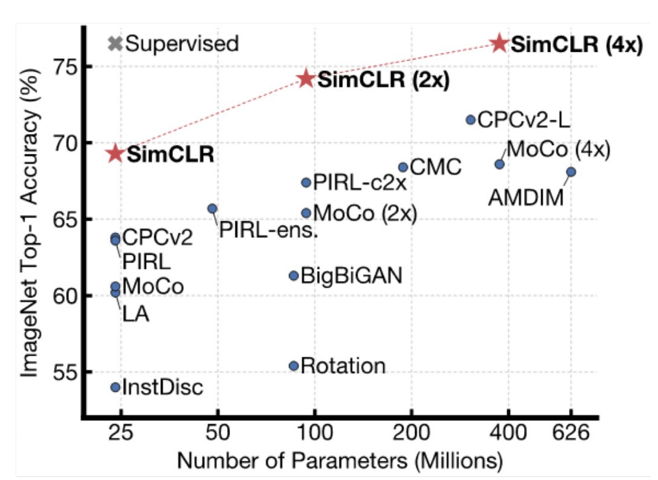


Learning Scheme

CURL aims to learn a representation function $\mathbf{f}$ by making **semantically similar (positive) pair closer** while **randomly drawn (negative) pair further**.

Inference

By performing fine-tuning linear classifier on top of the learned $\mathbf{f}$, we can get good empirical performance for the downstream task.



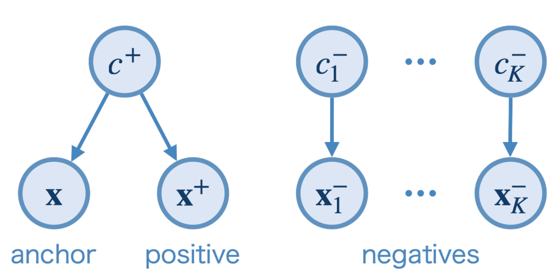
[Chen et al., 2020]

Q. What is the underlying mechanism of the success?

Problem Setup Based on [Arora et al., 2019]

Data Generating Process

- Draw 1 **positive**/K **negative** classes $c^+, \{c_k^-\}_{k\in[K]} \sim \mathbb{P}(Y)$
- Draw an anchor/positive sample $\mathbf{x}, \mathbf{x}^+ \sim \mathbb{P}(X|Y=c^+)$
- Draw K negative samples $\mathbf{x}_k^- \sim \mathbb{P}(X|Y=c_k^-)$



Training Objective

Train the representation function $\mathbf{f}$ by minimizing the following objective:

$$R_{\mathrm{cont}}(\mathbf{f}) = \mathop{\mathbb{E}}_{\substack{c^+,\{c_k^-\}\\ \mathbf{x},\mathbf{x}^+,\{\mathbf{x}_k^-\}}}\left[-\ln\frac{e^{\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)}}{e^{\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}^+)}+\sum_{k\in[K]}e^{\mathbf{f}(\mathbf{x})^\top\mathbf{f}(\mathbf{x}_k^-)}}\right]$$

Downstream performance

Evaluate the learned $\mathbf{f}$ by the downstream mean supervised loss:

$$R_{\mu\text{-supv}}(\mathbf{f}) = \mathop{\mathbb{E}}_{\mathbf{x},y\sim\mathbb{P}}\left[-\ln\frac{e^{\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_y}}{\sum_{c\in\mathcal{Y}}e^{\mathbf{f}(\mathbf{x})^\top\boldsymbol{\mu}_c}}\right]$$

Notations

C: # classes

K: # negative samples

$$\boldsymbol{\mu}_c = \mathbb{E}_{\mathbf{x}|c}[\mathbf{f}(\mathbf{x})] \qquad \mathbf{W}^\mu = [\boldsymbol{\mu}_1 \dots \boldsymbol{\mu}_C]^\top \qquad \inf_{\mathbf{W}\in\mathbb{R}^{C\times h}} R_{\mathrm{supv}}(\mathbf{W}\mathbf{f}) \le R_{\mu\text{-supv}}(\mathbf{f})$$

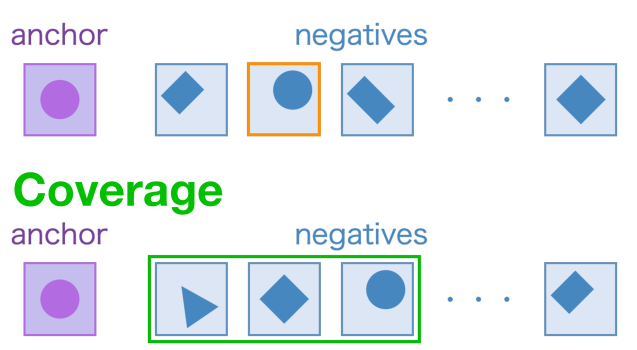
Analysis

Existing Work: Collision-Coverage Formulation

Collision-Coverage Formulation

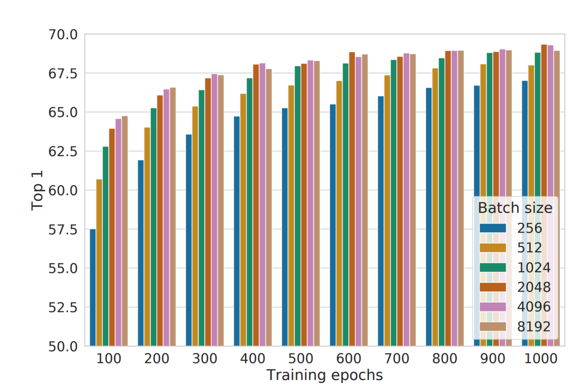
Rewrite the contrastive loss using the conditions under which the label **collision**/**coverage** occurs.

- **Collision**: randomly drawn negative class collides with the anchor class.
- **Coverage**: negative classes covers entire label space of the downstream classification.



Issue: Disagreement with Experiment

- Theory predicts the downstream performance degrades with increase in K because of the **label collision**, while larger K helps performance in practice.
- Upper bound becomes **exponentially loose** in K.



[Chen et al., 2020]

	Upper Bound	Reference
$R_{\mu\text{-supv}}(\mathbf{f}) \le$	$\frac{1}{(1-\tau_K)\upsilon_{K+1}}\{R_{\mathrm{cont}}(\mathbf{f}) - \mathbb{E}\ln(\mathrm{Col}+1)\}$	Arora et al. (2019)
	$\frac{1}{\upsilon_{K+1}}\{2R_{\mathrm{cont}}(\mathbf{f}) - \mathbb{E}\ln(\mathrm{Col}+1)\}$	Nozawa & Sato (2021)
	$\frac{2}{1-\tau_K}\left\lceil\frac{2(C-1)H_{C-1}}{K}\right\rceil\{R_{\mathrm{cont}}(\mathbf{f}) - \mathbb{E}\ln(\mathrm{Col}+1)\}$	Ash et al. (2022)

τ_K : collision prob.

υ_{K+1}: coverage prob.

Our Approach: Surrogate Bound

Main Result

Directly transform the contrastive loss to the supervised loss by **linearizing the log-sum-exp functions**.

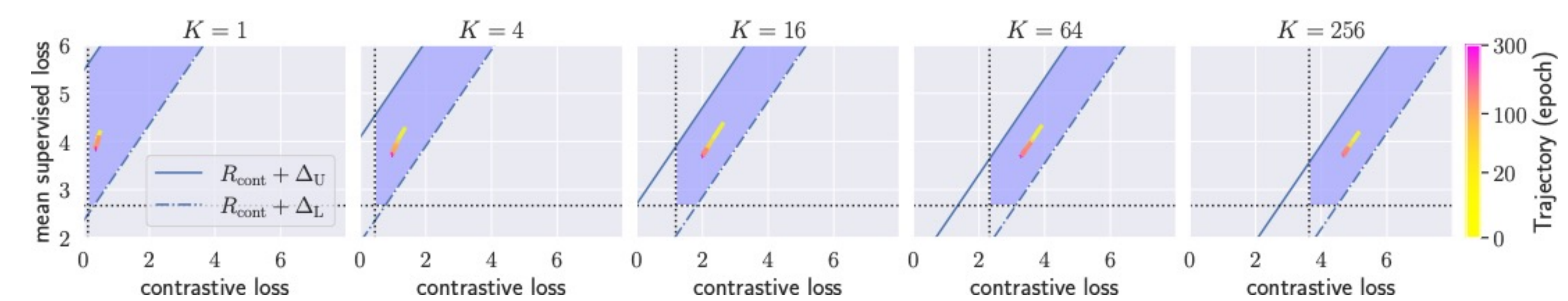
$$R_{\mathrm{cont}}(\mathbf{f}) + \Delta_L \le R_{\mu\text{-supv}}(\mathbf{f}) \le R_{\mathrm{cont}}(\mathbf{f}) + \Delta_U$$

$$\Delta_L = \Delta_U = O\left(\ln\frac{1}{K}\right)$$

- We can interpret the contrastive loss as the surrogate estimator of the mean supervised loss in a sense that these two losses behave similarly.
- Coefficients of the bounds are constant with respect to C and K.
- Surrogate gap (intercept) decreases as K increases; agrees with experimental facts.

Experiments

Synthetic Dataset



Setting

- 2D synthetic dataset `circle` with $C = 10$.
- $\mathbf{f}$: 3-layer MLP (# hidden units is 256) with ReLU activation.

Result

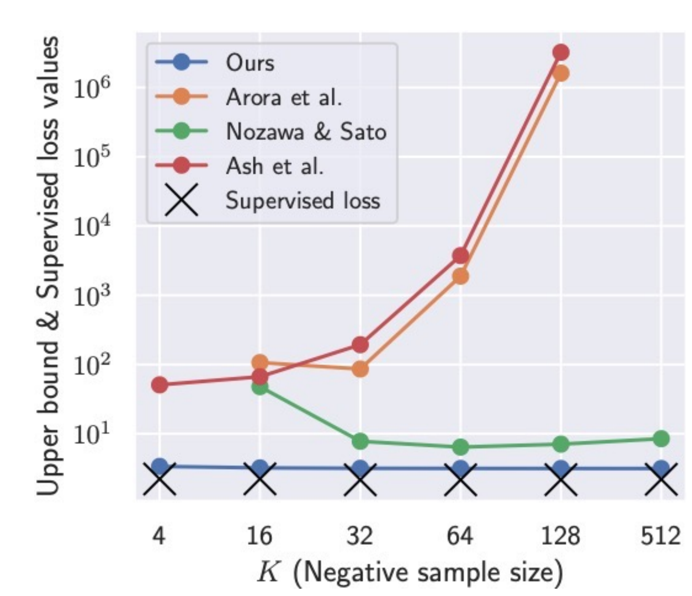
- Our surrogate bounds capture the learning dynamics well in different negative sample sizes.

Vision & Language Datasets

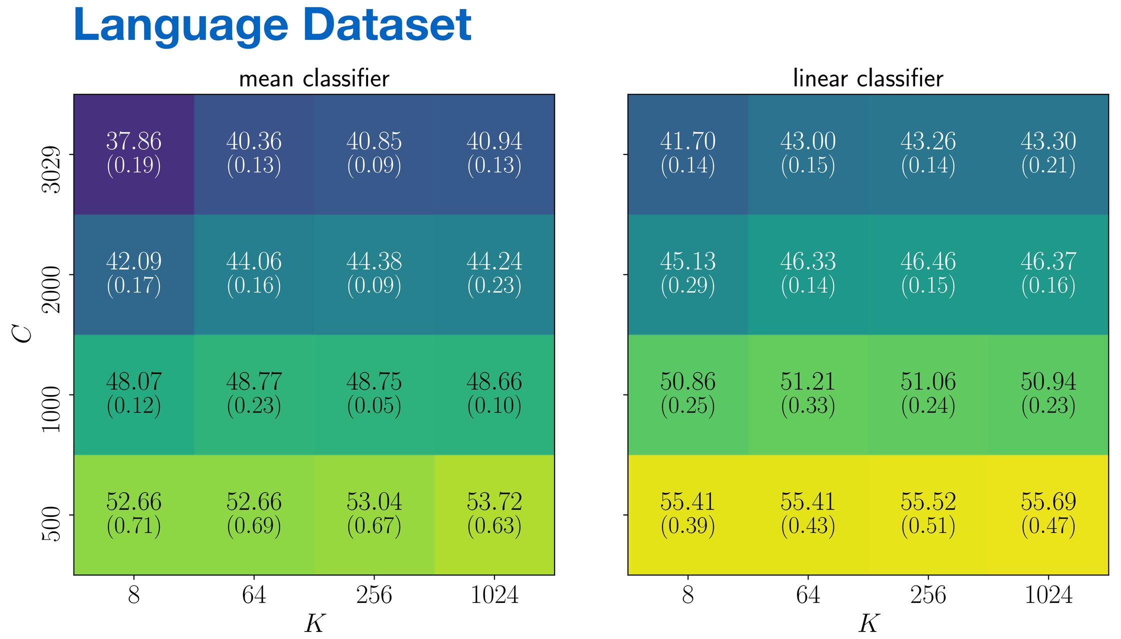
Setting

- Dataset: CIFAR-10/100 (vision) & Wiki-3029 (language).
- $\mathbf{f}$: ResNet-18-based [He et al., 2016] (vision) & fasttext-based [Joulin et al., 2017] (language)

Vision Dataset



Language Dataset



- Existing theories result in exponentially loose prediction of the downstream supervised loss for the test data in the vision dataset.
- Proposed upper bound agrees with the actual supervised loss well in all range of K.
- Larger K moderately helps performance as predicted from our theory.

A. Contrastive loss behaves as the surrogate estimator.