A Theoretical Analysis of Contrastive Unsupervised Representation Learning

Sanjeev Arora, Hrishikesh Khandeparkar, Mikhail Khodak, Orestis Plevrakis, Nikunj Saunshi

{arora,hrk,orestisp,nsaunshi}@cs.princeton.edu, khodak@cmu.edu

Mysterious Success of Contrastive Learning

Unsupervised methods for representation learning, reminiscent of word2vec for word embeddings, have been very successful in NLP [1] and to some extent in vision [2]. With access to **semantically similar points** and **random negative samples** from **unlabeled data**, they minimize objectives that look like

$$L_{unsuperv.}(f) = \mathop{\mathbb{E}}_{x,x^+,x^-}\left[\log\left(1+e^{f(x)^Tf(x^-)-f(x)^Tf(x^+)}\right)\right]$$

Why are these representations successful on future **linear classification tasks**? We attempt to demystify this by providing

- **Framework** connecting unlabeled data with downstream tasks
- **Provable guarantees** for such algorithms under the framework: Unsupervised loss is **surrogate** for *average supervised loss*

Framework

Semantic similarity ≈ membership in **same latent class**.

Connection

$\mathcal{X}$: Set of inputs, $\mathcal{C}$: Set of classes, ρ: Distribution over $\mathcal{C}$

$\mathcal{D}_c$: Universal distribution over $\mathcal{X}$ conditioned on class c.

Unlabeled Data

Similarity data: $(x,x^+)\sim\mathcal{D}_{sim}$

$c^+\sim\rho$

$(x,x^+)\sim\mathcal{D}_{c^+}^2$

Negative samples: $x^-\sim\mathcal{D}_{neg}$

$c^-\sim\rho$

$x^-\sim\mathcal{D}_{c^-}$

Supervised Tasks

Task: Subset of latent classes

$\mathcal{T}=\{c_1,\dots,c_k\}\subseteq\mathcal{C}$

Labeled samples: $(x,c)\sim\mathcal{D}_\mathcal{T}$

$c\sim\mathcal{T}$

$x\sim\mathcal{D}_c$

Evaluation Metric (Binary)

$$L_{sup}(\{c_1,c_2\},f) = \min_{\|w\|\le R}\ \tfrac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{c_1}}\log\left(1+e^{-f(x)^Tw}\right) + \tfrac{1}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{c_2}}\log\left(1+e^{f(x)^Tw}\right)$$

$$L_{sup}(f) = \mathop{\mathbb{E}}_{(c_1,c_2)\sim\rho^2}\left[L_{sup}(\{c_1,c_2\},f)\mid c_1\neq c_2\right]$$

Unsupervised Loss Bounds Supervised Loss

$\mathcal{F}\subseteq\{f:\mathcal{X}\to\mathbb{R}^d,\ \|f(\cdot)\|\le R\}$: Function class of interest.

τ: Probability that two classes sampled from ρ are the same.

$\hat{f}$: Minimizer from $\mathcal{F}$ of **empirical unsupervised loss**.

$$L_{unsup}(f) = \mathop{\mathbb{E}}_{\substack{(x,x^+)\sim D_{sim}\\ x^-\sim D_{neg}}}\left[\log\left(1+e^{f(x)^Tf(x^-)-f(x)^Tf(x^+)}\right)\right]$$

$\boldsymbol{L^\mu_{sup}(f)}$ is defined as loss of f when the **difference of means** classifier $w=\mu_{c_1}-\mu_{c_2}$ is used for the task $T=\{c_1,c_2\}$, where $\mu_c=\mathop{\mathbb{E}}_{x\sim\mathcal{D}_c}[f(x)]$. Clearly $L_{sup}(f)\le L^\mu_{sup}(f)$.

Key observation: *Jensen's inequality* to upper bound supervised loss. **Mean is better than random point** as classifier.

$$\log\left(1+e^{f(x)^T(\mu_{c^-}-\mu_{c^+})}\right) \le \mathop{\mathbb{E}}_{\substack{x^+\sim\mathcal{D}_{c^+}\\ x^-\sim\mathcal{D}_{c^-}}}\log\left(1+e^{f(x)^Tf(x^-)-f(x)^Tf(x^+)}\right)$$

Price of Negative Sampling: Class Collision

Inherent limitation of contrastive learning: negative samples can be from same class as similar pair $\implies$ $L_{un}(f)$ can be large. Need to understand when L_{un} can be made small

$$L_{un}(f)-\tau = \underbrace{(1-\tau)\,L^{\neq}_{un}(f)}_{c^+\neq c^-} + \underbrace{\tau\,(L^{=}_{un}(f)-1)}_{c^+=c^-}$$

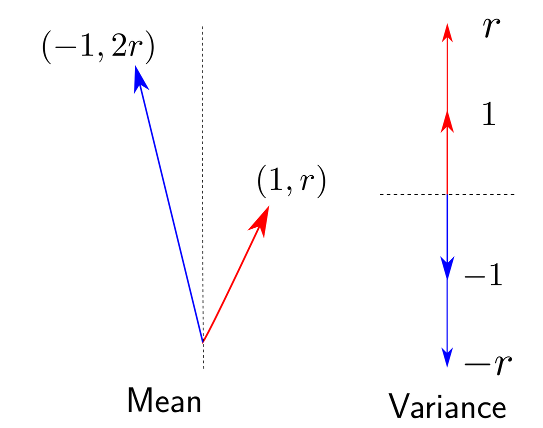
need contrastive f — **need intraclass concentration**

Theorem 2: Sufficient Conditions on $\mathcal{F}$

Let $x\sim\mathcal{D}_c$. λ_c: maximum standard deviation of $f(x)$ in a direction, R_c: mean norm of $f(x)$. Let $s(f)=2\mathop{\mathbb{E}}_{c\sim\rho}\lambda_cR_c$,

$$L_{sup}(\hat{f}) \le \boldsymbol{L^{\neq}_{un}(f)} + \frac{1}{1-\tau}\left[\tau\boldsymbol{s(f)}+Gen_M\right]\ ,\forall f\in\mathcal{F}$$

Stronger Competitive Guarantees?



$$L_{sup}(\hat{f}) \le L_{sup}(f)+Gen_M,\ \forall f \quad ✗$$

$$L^\mu_{sup}(\hat{f}) \le L^\mu_{sup}(f)+Gen_M,\ \forall f \quad ✗$$

Competitive bound: Need high-margin mean classifier and strong intraclass concentration. $L^\mu_{\gamma,sup}$ uses hinge loss with margin γ.

Lemma: Subgaussian Classes

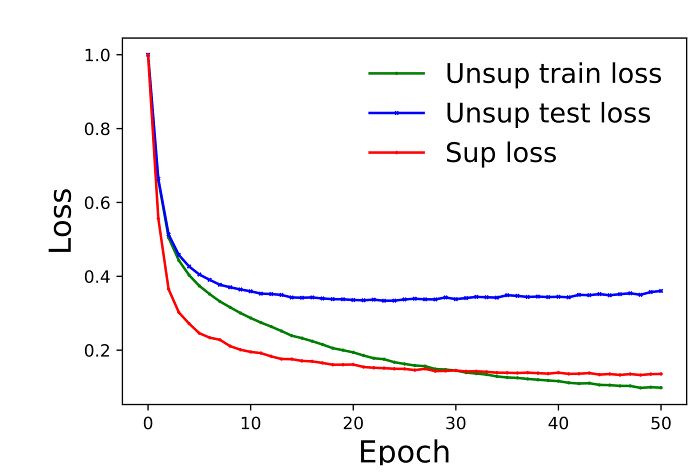
If f is σ^2-subgaussian within each class and $\gamma=1+\tilde{\Omega}(\sigma R)$

$$L^\mu_{1,sup}(\hat{f}) \le \boldsymbol{\gamma L^\mu_{\gamma,sup}(f)} + \frac{1}{1-\tau}\left[\tau s(f)+Gen_M\right]$$

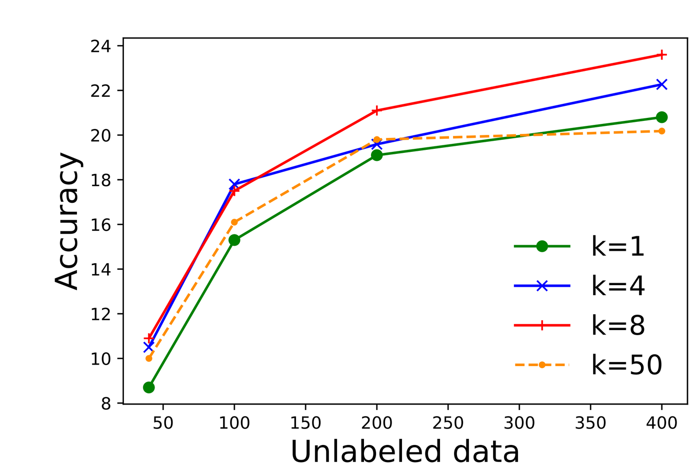
Extensions and Experiments

Multiple Negative Samples: Can bound the loss of average $(k+1)$-wise task with k negative samples. Increasing negative samples can hurt beyond a point (increased class collision).

Blocks of Similar Data: Can use the mean within a block as a proxy classifier. Gets tighter upper bound and improves performance on IMDb classification (beating SOTA model in [1]).



(a) Supervised loss roughly tracks unsupervised test loss as predicted by the Theorem 1



(b) Effect of amount of unlabeled data M and number of negative samples k. Very large k hurts.

		Supervised Tr	Supervised μ	Unsupervised Tr	Unsupervised μ
Wiki-3029	avg-2	97.8	97.7	97.3	97.7
	top-10	67.4	59.0	64.7	59.0
	top-1	43.2	33.2	38.7	30.4
CIFAR-100	avg-2	97.2	95.9	93.2	92.0
	top-5	88.9	83.5	70.4	65.6
	top-1	72.1	69.9	36.9	31.8

Table: Performance of supervised and unsupervised representations on average k-wise classification tasks (AVG-k) and full multiclass TOP-1 (not covered by theory). Classifier can be trained (Tr), or the mean is used (μ).

[1] Logeswaran & Lee. *An Efficient Framework for Learning Sentence Representations.* ICLR 2018.

[2] Wang & Gupta. *Unsupervised Learning of Visual Representations Using Videos.* ICCV 2015.