Stochastic Contrastive Learning

Jason Ramapuram*, Dan Busbridge*, Xavier Suau, Russ Webb
2nd Workshop on Self-Supervised Learning: Theory and Practice (NeurIPS 2021) · Apple Inc.

Abstract

Self-Supervised Learning (SSL) models:
- Present competitive performance with Supervised Learning,
- Lack the ability to infer latent variables.

This work (StochCon) introduces latent variables into the SimCLR contrastive learning framework and enables:
- Attributing representation uncertainty,
- Task-specific compression,
- Interpretable representations.

Objectives

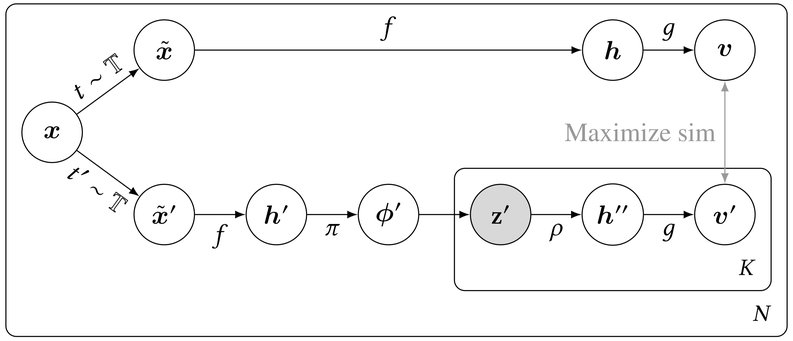
$$\mathscr{L}^{(i,j)}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(v_i, v_j)/\tau)}{\sum_{k=1}^{2N} 1_{[k \neq i]} \exp(\text{sim}(v_i, v_k)/\tau)}.$$

- Optimize standard InfoNCE objective.
- Representation vector z' is sampled from a pathwise differentiable distribution.

Contributions

- Introduce **differentiable latent variables** into SimCLR framework.
- StochCon-Bern induces a **588x compressed representation** of image data that is useful for downstream tasks.
- **Improves fine-tuned downstream performance** on CIFAR10 and ImageNet using ResNet50 and ResNet200.
- Demonstrate **competitive discriminative performance** on CIFAR10 with **as few as 11 bits**.

StochCon Model



Algorithm 1 Stochastic Contrastive Learning (StochCon)

Require: Data: $x \sim p(x), t \sim \mathbb{T}(x)$
Require: Models: f_θ : backbone, g_θ : head, $\{\pi_\theta, \rho_\theta\}$: projectors
while not converged **do**
- $\{\tilde{x}, \tilde{x}'\} = \{t \circ x, t' \circ x\}$ ▷ Augment input with $\{t, t'\}$
- $\{h, h'\} = \{f_\theta(\tilde{x}), f_\theta(\tilde{x}')\}$ ▷ Produce representations
- $\phi' = \pi_\theta(h')$ ▷ (optional) Bottleneck projection
- $z' \sim q_\theta(z|x)$ ▷ Pathwise differentiable (Mohamed et al., 2020) latent variable.
- $h'' = \rho_\theta(z')$ ▷ (optional) Bottleneck upsampler
- $\{v, v'\} = \{g_\theta(h), g_\theta(h'')\}$ ▷ InfoNCE projection
- $\min_\theta \mathcal{L}_{\text{InfoNCE}}(v, v')$

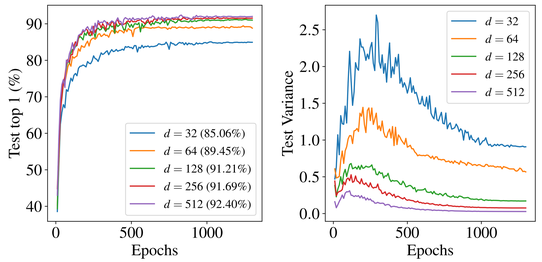
end while

Results

Test Top-1 %

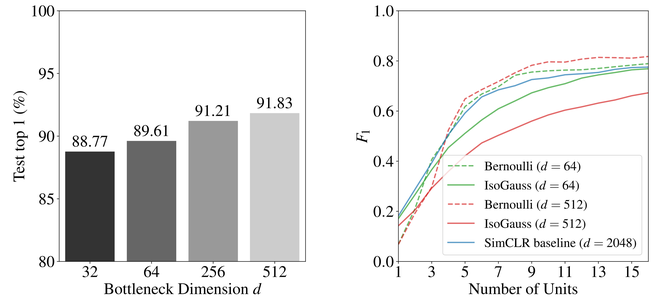
	CIFAR10-ResNet50		ImageNet-ResNet50		ImageNet-ResNet200	
Model	Fine-Tuned	Frozen	Fine-Tuned	Frozen	Fine-Tuned	Frozen
StochCon Bern	**96.42**	91.96	**77.49**	67.00	**80.24**	64.25
StochCon Iso-Gauss	96.08	**92.40**	—	—	—	—
Supervised	95.00	—	76.13	—	78.34	—
SimCLR	94.35	91.35	76.37	**71.34**	79.82	**73.52**

StochCon CIFAR10 Isotropic-Gaussian Ablations

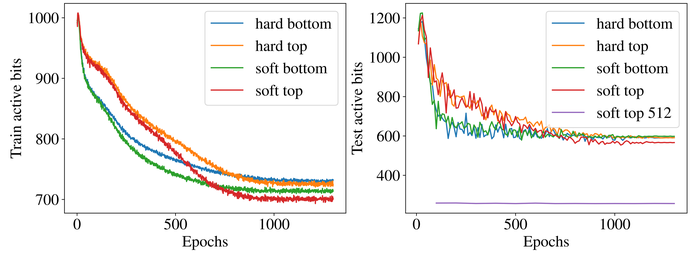


- Prevent variance collapse by estimating variance of opposing set of views.
- Smaller bottleneck dimension relies on variance to increase representation capacity.

StochCon CIFAR10 Bernoulli Ablations



- Bottleneck trades top-1 performance for compression.
- Mean F1 performance of multi-class Random Forest with varying number of features $\mapsto$ as few as 11 features leads to competitive CIFAR10 performance.



- StochCon-Bern enables countable metrics.
- Bottleneck representation (soft top 512) uses all available capacity (50% zeros, 50% ones).