EMC²: Efficient MCMC Negative Sampling for Contrastive Learning with Global Convergence

Chung-Yiu Yau[1], Hoi-To Wai[1], Parameswaran Raman[2], Soumajyoti Sarkar[2], Mingyi Hong[3]
[1]The Chinese University of Hong Kong [2]Amazon Web Services [3]University of Minnesota

Summary / TL;DR

Existing contrastive learning algorithms either
1) require large batch for high accuracy, or
2) is unstable.

We propose **EMC²** that
1) converges to high accuracy with **small batch**,
2) with theoretical convergence of $\mathcal{O}(1/\sqrt{T})$.

Contrastive Learning - Definition

Contrastive learning finds the feature encoders $\phi^\star, \psi^\star$ that maximizes similarity $\phi^\star(x)^\top \psi^\star(y)$ between positive data pair (x, y) and minimizes similarity $\phi^\star(x)^\top \psi^\star(z)$ between negative data pair (x, z).

Contrastive Loss Function

❖ InfoNCE loss (e.g., CLIP [1], SimCLR [2]) with mini-batch size B:

$$\mathcal{L}_{\rm NCE}(\theta; B) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\rm pos}} \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{D}_{\rm neg}(x;B)} \left[-\log \frac{\exp(\beta\, \phi(x;\theta)^\top \psi(y;\theta))}{\sum_{z\in\mathbf{Z}} \exp(\beta\, \phi(x;\theta)^\top \psi(z;\theta))} \right]$$

❖ Global contrastive loss (e.g., SogCLR [3]):

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\rm pos}} \left[-\log \frac{\exp(\beta\, \phi(x;\theta)^\top \psi(y;\theta))}{\sum_{z\in\mathbf{D}_{\rm neg}(x)} \exp(\beta\, \phi(x;\theta)^\top \psi(z;\theta))} \right]$$

❖ Global loss is the limiting upper bound of InfoNCE:

$$\mathcal{L}_{\rm NCE}(\theta; B) \le \mathcal{L}(\theta) \quad \forall B > 0$$

$$\lim_{B\to|\mathbf{D}_{\rm neg}|} \mathcal{L}_{\rm NCE}(\theta; B) = \mathcal{L}(\theta)$$

❖ We propose to minimize $\mathcal{L}(\theta)$, which upper bounds the large batch objective used in CLIP for **any batch size** $B > 0$, at the cost of **constant batch size** using MCMC sampling.

Global Loss Gradient

$$\nabla\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\rm pos}} \left[-\beta\, \nabla_\theta \big(\phi(x;\theta)^\top \psi(y;\theta)\big) \right] + \underbrace{\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_{\rm pos}} \left[\beta \sum_{z\in\mathbf{D}_{\rm neg}(x)} p_{x,\theta}(z)\, \nabla_\theta \big(\phi(x;\theta)^\top \psi(z;\theta)\big) \right]}_{\nabla\mathcal{L}_{\rm neg}(\theta)}$$

with a softmax distribution:

$$p_{x,\theta}(z) = \frac{\exp(\beta\, \phi(x;\theta)^\top \psi(z;\theta))}{\sum_{z'\in\mathbf{D}_{\rm neg}(x)} \exp(\beta\, \phi(x;\theta)^\top \psi(z';\theta))}$$

❖ Negative pair gradient $\nabla\mathcal{L}_{\rm neg}(\theta)$ admits a data-dependent softmax distribution $p_{x,\theta}(z)$.

EMC²: MCMC Sampling on $\nabla\mathcal{L}_{\rm neg}(\theta)$

❖ We propose to apply **Metropolis-Hasting** algorithm for sampling $\nabla\mathcal{L}_{\rm neg}(\theta)$.

❖ Accept a random negative sample Z_i' with probability

$$Q_{x_i,\theta}(Z_i', Z_i) = \frac{p_{x_i,\theta}(Z_i')}{p_{x_i,\theta}(Z_i)} = \frac{\exp(\beta\, \phi(x_i;\theta)^\top \psi(Z_i';\theta))}{\exp(\beta\, \phi(x_i;\theta)^\top \psi(Z_i;\theta))}$$

(**Hardness-aware** negative sampling)

❖ **$\mathcal{O}(B^2)$ Computation Overhead**: Only requires computing the acceptance probability $Q_{x_i,\theta}(Z_i', Z_i)$.

❖ **$\mathcal{O}(m)$ Memory Overhead**: Only requires storing the exponential score of previously accepted negative sample, for each x_i in the dataset of size m.

❖ **MCMC with Warm Starting**: Retain MC state from previous epoch and uses $\mathcal{O}(1)$ samples for each epoch, more efficient than $\mathcal{O}(1/\tau_{\rm mix})$ samples in Cold Started MCMC.

❖ **Convergence**: We guaranteed EMC² converges at the rate of $\mathcal{O}(1/\sqrt{T})$.

Experiments

❖ EMC² shows competitive **small batch performance**.

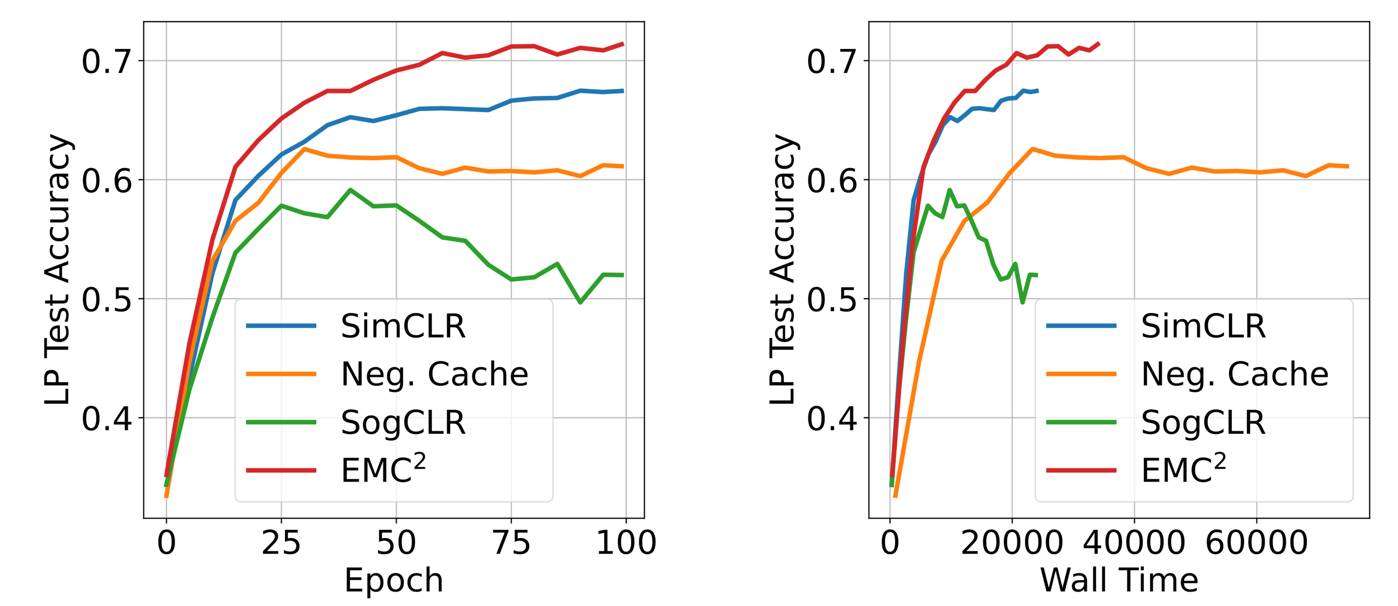

Figure 1: Training ResNet-18 on STL-10 using Adam with batch size $b = 32$, compared on linear probe accuracy.

❖ EMC² **converges accurately** with batch size $b = 4$.

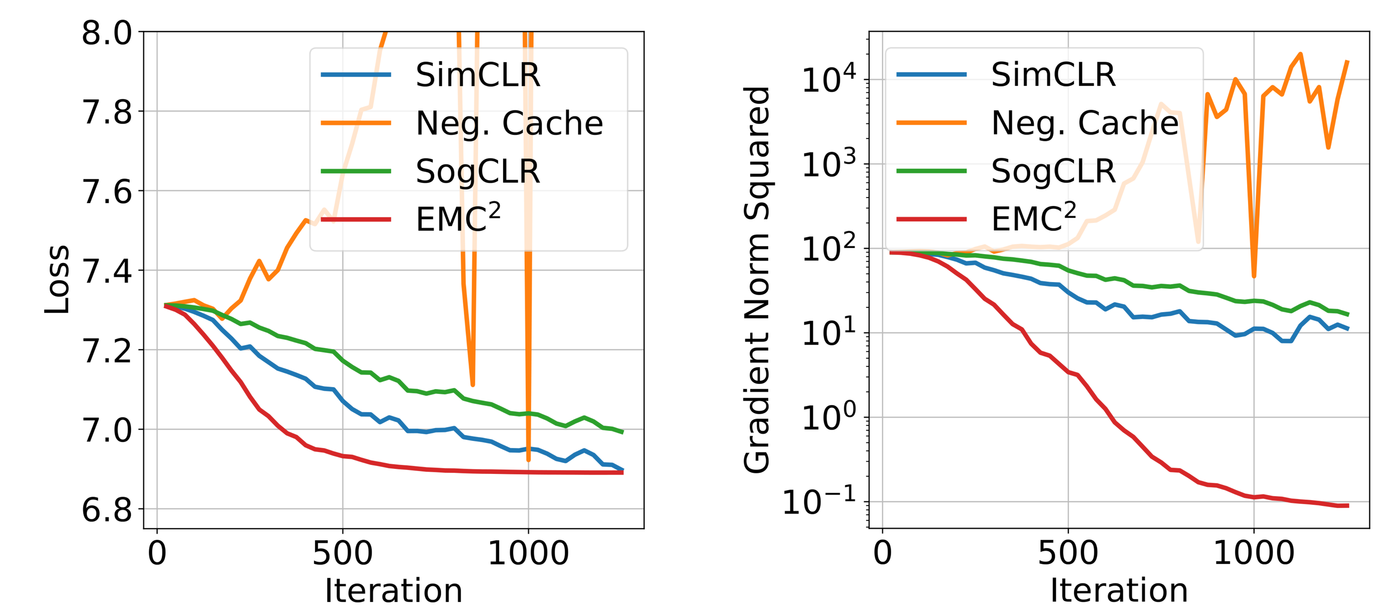

Figure 2: Comparison on a subset of STL-10 using the first 500 images and pre-computed two augmentations for each image. Trained using SGD with batch size $b = 4$.

References

[1] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. *In International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.
[2] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. *In International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020.
[3] Yuan, Z., Wu, Y., Qiu, Z.-H., Du, X., Zhang, L., Zhou, D., and Yang, T. Provable stochastic optimization for global contrastive learning: Small batch does not harm performance. *In International Conference on Machine Learning*, pp. 25760–25782. PMLR, 2022.

ICML 2024 Poster Session @ Messe Wien Exhibition Congress Center, Vienna, Austria

*Corresponding author: oscarcyy@amazon.com, cyyau@se.cuhk.edu.hk