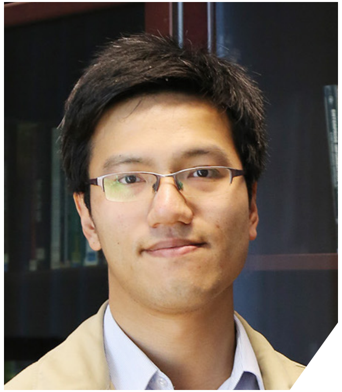
EMC^2: Efficient MCMC Negative Sampling for Contrastive Learning with Global Convergence

July 15, 2024

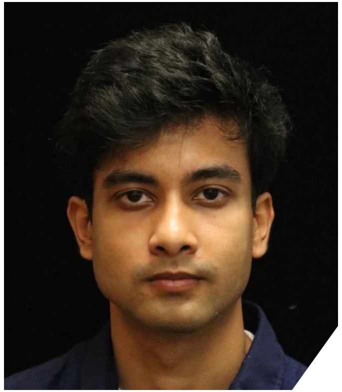
Authors

Chung-Yiu Yau
AWS Applied Scientist Intern, PhD Candidate @ The Chinese University of Hong Kong

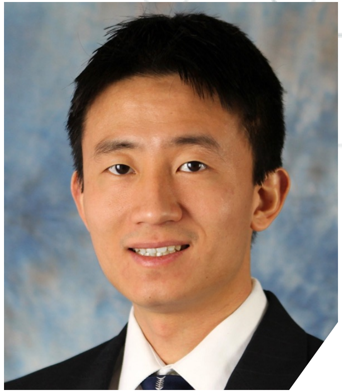


Hoi-To Wai
Professor, The Chinese University of Hong Kong

Parameswaran Raman
AWS Applied Scientist



Soumajyoti Sarkar
AWS Applied Scientist



Mingyi Hong
Professor, University of Minnesota

amazon | science

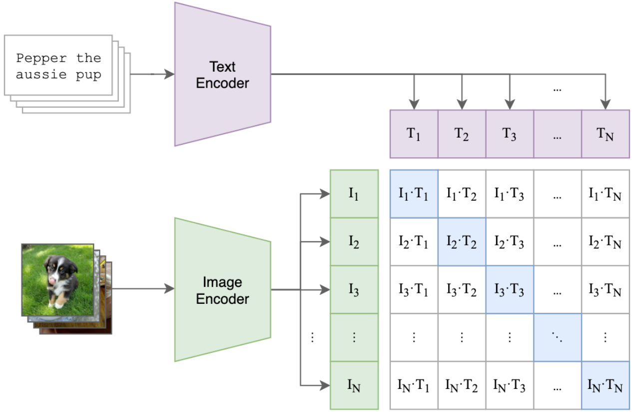


Image Source: Radford et. al., Learning transferable visual models from natural language supervision, ICML, 2021.

Contrastive Learning

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 with mini-batch size B

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

(adopted in CLIP [1], SimCLR [2])

Global Contrastive Loss

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

(adopted in SogCLR [3])

- Global contrastive loss is the limiting upper bound of InfoNCE loss as batch size increases:

$$\mathcal{L}_{\text{NCE}}(\theta; B) \le \mathcal{L}(\theta) \quad \forall B > 0 \qquad\qquad \lim_{B\to|\mathbf{D}_{\text{neg}}|} \mathcal{L}_{\text{NCE}}(\theta; B) = \mathcal{L}(\theta)$$

[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.

amazon | science

We propose to minimize the global contrastive loss $\mathcal{L}(\theta)$, 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.

amazon | science

Global Loss Gradient

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

$$\equiv \nabla\mathcal{L}_{\text{pos}}(\theta) + \nabla\mathcal{L}_{\text{neg}}(\theta)$$

with a softmax distribution:

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

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

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

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

- Overhead due to Metropolis-Hasting Sampling:
  - $\mathcal{O}(\boldsymbol{B^2})$ **Computation Overhead**: Only requires computing the acceptance probability $Q_{x_i,\theta}(Z_i', Z_i)$.
  - $\mathcal{O}(\boldsymbol{m})$ **Memory Overhead**: Only requires storing the exponential score $\exp(\beta\, \phi(x_i;\theta)^\top \psi(Z_i;\theta))$ of previously accepted negative sample Z_i, for each x_i in the dataset of size m.
- **MCMC with Warm Starting**: Retain Markov Chain state Z_i from previous epoch and uses $\mathcal{O}(1)$ samples for each epoch, more efficient than $\mathcal{O}(1/\tau_{\text{mix}})$ 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**.

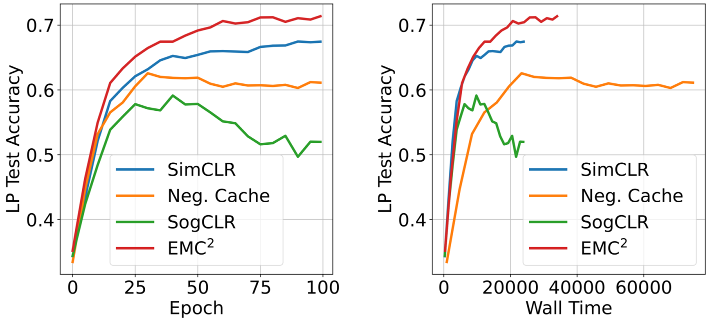


Figure 1: Training ResNet-18 on STL-10 using Adam with batch size $b = 32$, compared on linear probe accuracy.

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

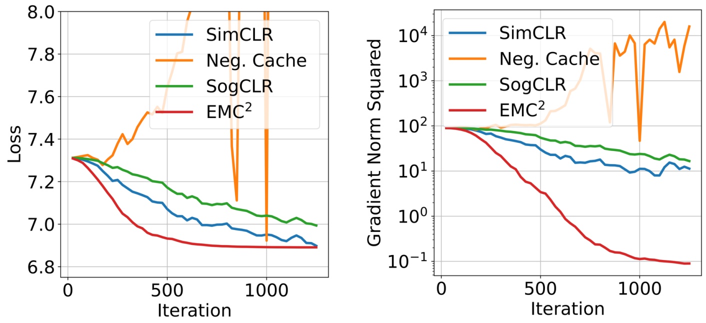


Figure 2: Comparison on a subset of STL-10 using the first 500 images and pre-computed two augmentations for each image. Trained using SGD with batch size $b = 4$.

Thank You.

Paper

Github

amazon | science