Google DeepMind

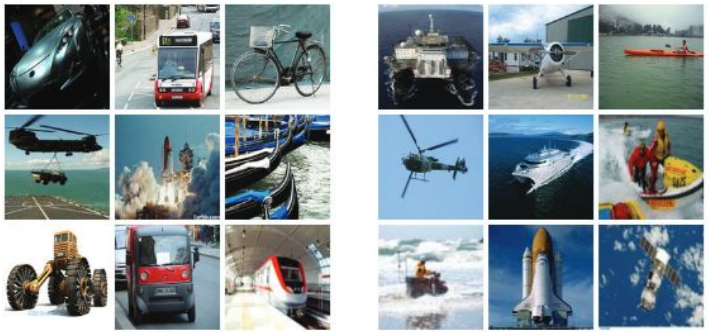
Representation Learning via Non-Contrastive Mutual Information

Zhaohan Daniel Guo
Bernardo Avila Pires
Khimya Khetarpal
Dale Schuurmans
Bo Dai

Poster #169

Self-Supervised Representation Learning

Lots of unlabeled data?



Labeling too expensive?

Self-Supervised Representation Learning

Lots of unlabeled data?

Labeling too expensive?

Self-Supervised Representation Learning!

Contrastive vs. Non-Contrastive

Contrastive

eg) SimCLR, Spectral Contrastive, f-MICL

PROS

Theoretically Sound

Maximize Mutual Information

CONS

High Variance Loss

Pairwise comparison of all data points

Non-Contrastive

eg) BYOL, SimSiam

PROS

Batch Size Efficient

CONS

Potential Representation Collapse

Theory-Practice Gap

Contrastive vs. Non-Contrastive

Contrastive

eg) SimCLR, Spectral Contrastive, f-MICL

PROS

Theoretically Sound

Maximize Mutual Information

CONS

High Variance Loss

Pairwise comparison of all data points

Non-Contrastive

eg) BYOL, SimSiam

PROS

Batch Size Efficient

CONS

Potential Representation Collapse

Theory-Practice Gap

MINC Loss - Best of Both Worlds!

MINC Loss - Step 1

Start with Spectral Contrastive Loss

Theoretically maximizes a certain mutual information.

$$\max_{\phi} \underbrace{2\mathbb{E}_{p(x,x')}\left[\phi(x)^\top \phi(x')\right]}_{\text{Align embeddings for similar data points.}} - \underbrace{\mathbb{E}_{p(x)p(x')}\left[\left(\phi(x)^\top \phi(x')\right)^2\right]}_{\text{Make dissimilar embeddings orthogonal}} \tag{3}$$

MINC Loss - Step 2

Refactor with Auxiliary

Avoids double expectation.

$$\max_{\phi} \; 2\mathbb{E}_{p(x,x')}\left[\phi(x)^\top \phi(x')\right] - \mathbb{E}_{p(x)p(x')}\left[\left(\phi(x)^\top \phi(x')\right)^2\right] \tag{3}$$

$$\max_{\phi} \; 2\mathbb{E}_{p(x,x')}\left[\phi\left(x\right)^\top \phi\left(x'\right)\right] - \mathbb{E}_{p(x')}\left[\phi\left(x'\right)^\top \Lambda \phi\left(x'\right)\right]$$

$$\text{subject to} \quad \Lambda = \mathbb{E}_{p(x)}\left[\phi\left(x\right)\phi^\top\left(x\right)\right], \tag{5}$$

Given Lambda (d x d matrix), this becomes a non-contrastive loss!

MINC Loss - Step 3

Use EMA and Generalized Hebbian Algorithm (GHA)

EMA Update: $\Lambda_{t+1} \leftarrow (1-\alpha)\Lambda_t + \alpha(\phi(x)\phi(x)^\top)$

$$\max_{\phi} \; 2\mathbb{E}[\phi(x)^\top \phi(x)] + \mathbb{E}[\phi(x)^\top \mathbf{LT}(\Lambda)\phi(x)]$$

Lower Triangular Transformation:

From Generalized Hebbian Algorithm - Asymptotic Orthogonality of Embeddings

Approximate Power Iteration

Framing MINC as Approximate Power Iteration

The MINC objective can be seen doing approximate power iteration for finding the eigendecomposition of a low-rank approximation:

Low-Rank Eigendecomposition: $\min_{\phi} \|M - FF'^{\top}\|_{\mathrm{F}}^2,$

Iterate Fixed Point Equation: $\int \frac{p(x,x')}{\sqrt{p(x)p(x')}} \sqrt{p(x)}\phi(x)\,dx = \Lambda\sqrt{p(x')}\phi(x')$

Subject to Orthogonality: $\int \phi(x)\,\phi(x)^{\top} p(x)dx = \int \phi(x')\,\phi(x')^{\top} p(x')dx' = \Lambda,$

Approximate Power Iteration

Framing MINC as Approximate Power Iteration

The MINC objective can be seen doing approximate power iteration for finding the eigendecomposition of a low-rank approximation:

Low-Rank Eigendecomposition: $\min_{\phi} \|M - FF'^{\top}\|_{\mathrm{F}}^2,$

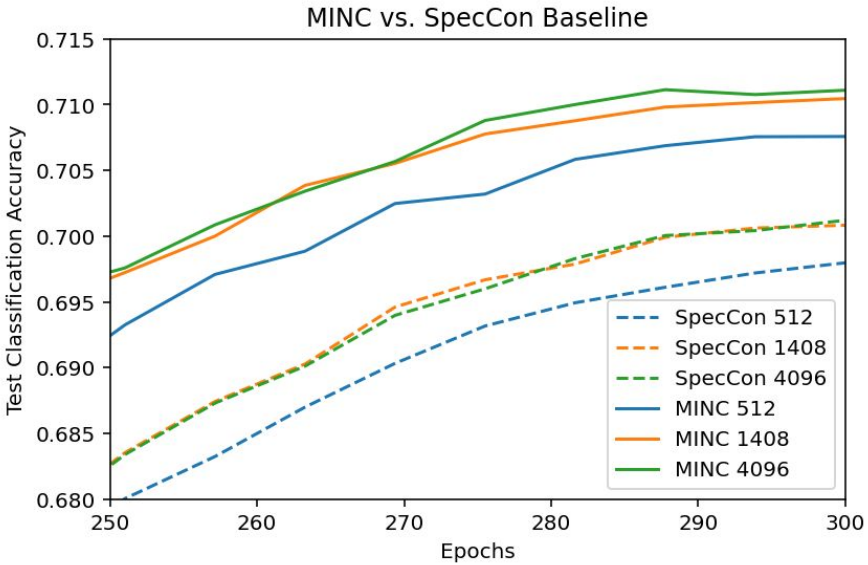
Iterate Fixed Point Equation: $\int \frac{p(x,x')}{\sqrt{p(x)p(x')}} \sqrt{p(x)}\phi(x)\, dx = \Lambda \nu$

Subject to Orthogonality: $\int \phi(x)\,\phi(x)^{\top} p(x) dx = \int \phi(x')\,\phi(x')^{\top} p(x') dx' = \Lambda$

Experimental Results

MINC Outperforms Spectral Contrastive Loss at multiple batch sizes!

MINC at small batch size of 512 outperforms Spectral Contrastive at large batch size of 4096! More results in poster #169 and paper!



Thank you!
See Poster #169 today (May 3)