CSE6243: Advanced Machine Learning **Fall 2024**

Lecture 19: Representation Learning from Spectral Decomposition view

Lecturer: Bo Dai *Scribes: Vidhya Kewale & Sandilya Sai Garimella*

Note: *LaTeX template courtesy of UC Berkeley EECS Department.*

19.1 Recap

SimCLR (**Sim**ple **C**ontrastive **L**earning of Visual **R**epresentations)

- self-supervised contrastive learning method for visual representations
- works with a single modality: images
- trains the model to bring augmented views of the same image closer together in the embedding space, while pushing apart representations of different images

CLIP (**C**ontrastive **L**anguage-**I**mage **P**re-training)

- extends contrastive learning to multiple modalities using image-text pairs
- learns aligned representations across the visual and textual domains
- trains the model to maximize the cosine similarity between embeddings of matching image-text pairs, while minimizing it for non-matching pairs

Energy-Based Models (EBM) and Noise-Contrastive Estimation (NCE)

1. conditional probability in the same modality

$$p(x' \mid x) = p(x') \exp(\varphi(x)^\top \varphi(x^\top)) \tag{1.1}$$

2. cross-modal conditional probabilities

$$p(y \mid x) = p(y) \exp(\varphi(x)^\top \nu(y)) \tag{1.2}$$

$$p(x \mid y) = p(x) \exp(\varphi(x)^\top \nu(y)) \tag{1.3}$$

19.2 SimCLR

All of them use ranking-based NCE to estimate a special EBM. SimCLR is as follows:

$$p\left(x' \mid x\right)=p\left(x'\right) \exp \left(\varphi(x)^{\top} \varphi\left(x'\right)\right) \tag{2.1}$$

$$D=\{(x_i, x_i', (x_i'^1, ..., x_i'^k))\}_{i=1}^n \tag{2.2}$$

We formulate the loss function using the follows:

$$\max_{\varphi_\theta} \sum_{i=1}^{n}\left[\varphi\left(x_{i}\right)^{\top} \varphi\left(x_{i}'\right)-\log \sum_{j=1}^{k} \exp \left(\varphi\left(x_{i}\right)^{\top} \varphi\left(x_{i}'^{j}\right)\right)\right] \tag{2.3}$$

$$\max_{\varphi} f(\theta); \quad D_\theta l(\theta)=\sum_{i=1}^{n} \varphi\left(x_{i}\right)^{\top} \varphi\left(x_{i}'\right) \cdot\left(\nabla_{\theta} \varphi\left(x_{i}\right)+\nabla_{\theta} \varphi\left(x_{i}'\right)\right) \tag{2.4}$$

$$=\sum_{i=1}^{n}\left(\sum_{j=1}^{k} \frac{\exp \left(\varphi\left(x_{i}\right)^{\top} \varphi\left(x_{i}'^{j}\right)\right) \cdot\left(\nabla_{\theta} \varphi\left(x_{i}\right)+\nabla_{\theta} \varphi\left(x_{i}'^{j}\right)\right)}{\sum_{j=1}^{k} \exp \left(\varphi\left(x_{i}\right)^{\top} \varphi\left(x_{i}'^{j}\right)\right)}\right)=\mathcal{O}(nk) \tag{2.5}$$

Computation cost is $\mathcal{O}(nk)$. We typically also use k=n, therefore the computation cost becomes $\mathcal{O}(n^2)$.

To further demonstrate with data, assume we have:

$$\{x_i\}_{i=1}^B \sim \{x_i'\}_{i=1}^B \tag{2.6}$$

Therefore, the computation cost will be:

$$i, \{x-i\} \quad (B-1) \Rightarrow \sim \mathcal{O}\left(B^2\right) \tag{2.7}$$

To circumvent this quadratic computation cost, we can use a binary-based NCE instead of a ranking-based NCE. With this, instead of $\mathcal{O}(nk)$, we can get $\mathcal{O}(2B) \sim \mathcal{O}(B)$.

Coming back to this expression, to derive spectral learning and Bootstrap your own latent (BYOL):

$$p\left(x^* \mid x\right)=p\left(x'\right) \exp \left(\varphi(x)^{\top} \varphi(x)\right) \tag{2.8}$$

We remove the exponential because it makes the gradient calculation harder:

$$p\left(x' \mid x\right)=p\left(x'\right) \varphi\left(x'\right)^{\top} \varphi(x) \tag{2.9}$$

The L2 loss function is now defined as:

$$l_2 \int \left\| p\left(x' \mid x\right) - p\left(x'\right) \varphi\left(x'\right)^{\top} \varphi(x) \right\|^2 dx dx' \tag{2.10}$$

$$= \int p\left(x' \mid x\right)^2 dx dx' - 2 \int p\left(x' \mid x\right) p\left(x'\right) \tag{2.11}$$

$$p\left(x'\right)^{\top} \varphi(x) dx dx' \tag{2.12}$$

We know $p\left(x' \mid x\right) p(x) = p\left(x'\right) p(x) p\left(x'\right)^{\top} p(x)$ from $p\left(x' \mid x\right) = p\left(x'\right) \varphi\left(x'\right)^{\top} \varphi(x)$:

$$\int \left\| \frac{p\left(x', x\right)}{\sqrt{p(x)}\sqrt{p(x)}} \sqrt{p\left(x'\right)}\sqrt{p(x)^2} \varphi\left(x'\right)' \varphi(x) \right\|^2 dx dx' \tag{2.13}$$

$$= \int \left(\frac{p\left(x', x\right)}{\sqrt{p(x)}\sqrt{p(x)}}\right)^2 dx dx' - 2 \int \left(p\left(x', x\right) p(x')^T \varphi(x)\right) dx dx' + \int p(x') p(x) (\varphi(x')^T \varphi(x))^2 dx dx' \tag{2.14}$$

We observe that the terms in the integrals can be simplified using the definition of expectation; therefore we can apply sampling here. The above simplifies to:

$$= -2\mathbb{E}_{p(x,x')} \left[\varphi\left(x'\right)^{\top} \varphi(x)\right] + \mathbb{E}_{p(xp(x)} \left[\left[\varphi(x')^{\top} (\varphi(x)\right)^2\right]. \tag{2.15}$$

From above, we can see that we sample only once but can use it for computing both expectation terms.

$$p\left(x', x\right) = p(x) \varphi(x)^{\top} p\left(x'\right) \varphi\left(x'\right) \tag{2.16}$$

but we write this as

$$p\left(x', x\right) = \Psi(x)^{\top} \Psi(x') \tag{2.17}$$

This is called the Eigen-decomposition spectral perspective of representation.

19.2.1 BYOL w/o ν

The loss function is, using similar reason to above:

$$\min_{\varphi,\nu} \int \left(\frac{\rho(x', x)}{\sqrt{\rho(x')}\sqrt{\rho(x)}} - \sqrt{\rho(x')}\sqrt{\rho(x)} \quad \nu(x')^{\top} \varphi(x) \right) ||^2 dx dx' \tag{2.18}$$

Alternative Optimization

Add a constraint such that $\nu = \varphi$.

$$\textbf{(min problem above)} \propto 2\mathbb{E}_{p(x',x)}\left[\nu(x)^T\varphi(x)\right] - \mathbb{E}_{p(x')}\left[\varphi(x')^T\mathbb{E}_{p(x)}\left[\varphi(x)\varphi(x)^\top\right]\varphi(x')\right] \tag{2.19}$$

With the above expanded, we can do separate sampling.

$$\Lambda_t = \mathbb{E}_{p(x)}\left[\nu_\Psi(x)\nu_\Psi(x)^T\right] \tag{2.20}$$

$$-2\mathbb{E}_{p(x,x')}\left[\varphi(x')\nu(x)^T\right] + \mathbb{E}_{p(x)}\left[\varphi(x)^T\Lambda_t\varphi(x)\right] \tag{2.21}$$

19.3 PCA

Finding the maximal eigenspace while matching the y's are different.

We have the following, noting that the trace operator is invariant under cyclic permutations:

$$\hat{\rho} = (x, x') \in \mathbb{R}^{n\times n}, n \text{ samples} \tag{3.1}$$

$$\Psi(x) \in \mathbb{R}^{n\times d} \tag{3.2}$$

$$\mathbb{E}_{p(x,x')}\left[\Psi(x)\Psi(x')^T\right] \tag{3.3}$$

$$\mathbb{E}_{p(x)}\left[\Psi(x)T\Psi(x)\right] = I_{d\times d} \tag{3.4}$$

Penalty method:

$$\max_\Psi \mathbb{E}_{p(x,x')}\left[\Psi(x)\Psi(x)^T\right] - \lambda\cdot\text{trace}(\mathbb{E}_{p(x)}\left(\Psi(x)\Psi(x)^T\right) - I)^2 \tag{3.5}$$