Revisiting Contrastive Learning as Spherical Sliced Wasserstein Maximization

Anonymous Submission

Abstract

Introduction

Proposed Model

Generalized contrastive learning

Given a set of samples $\mathcal{X}=\{x\}$, we would like to learn a representation model $f:\mathcal{X}\mapsto\mathcal{Z}$ in an unsupervised way and obtain the latent representations of the samples, $i.e.$, the $\mathcal{Z}=\{z\}$. The typical contrastive learning methods like noise-contrastive estimation (Gutmann and Hyvärinen 2010) achieve this aim by maximizing the difference between the conditional distribution of positive samples and that of negative samples.

$$\begin{aligned}\max\nolimits_f \mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{x'\sim p_{\mathcal{P}|x}}[s(f(x');f(x))]-\\ \mathbb{E}_{x'\sim p_{\mathcal{N}|x}}[h^*(s(f(x');f(x)))]],\end{aligned}\tag{1}$$

where $p_{\mathcal{X}}$ represents the (empirical) data distribution, $p_{\mathcal{P}|x}$ and $p_{\mathcal{N}|x}$ represent the positive and the negative sample distributions conditioned on the sample x. $s(\cdot;\cdot)$ is a score function of the latent representations $f(x')$, and the score often corresponds to the similarity between $f(x')$ and $f(x)$.

Following the work in (Nowozin, Cseke, and Tomioka 2016), we can explain the framework in (1) as maximizing the expectation of the f-divergence between $p_{\mathcal{P}|x}$ and $p_{\mathcal{N}|x}$ conditioned on various samples, where $h^*:\mathbb{R}\mapsto\mathbb{R}$ is a conjugate function, whose formulation is determined by the final activation layer of s. More specifically, when $s(f(x');f(x))=-\log(1+\exp(-f(x')^\top f(x)))$, $h^*(t)$ becomes $-\log(1-\exp(t))$, and we can rewrite (1) as a mutual information maximization problem (Hjelm et al. 2018; Veličković et al. 2018):

$$\begin{aligned}\max\nolimits_f \mathbb{E}_{x\sim p_{\mathcal{X}}}[\mathbb{E}_{x'\sim p_{\mathcal{P}|x}}[\log\sigma(f(x')^\top f(x))]+\\ \mathbb{E}_{x'\sim p_{\mathcal{N}|x}}[\log(1-\sigma(f(x')^\top f(x)))]],\end{aligned}\tag{2}$$

where σ is a sigmoid function.

In this work, we generalize (1) from a different view point. Essentially, the optimization problem in (1) aims at leveraging a (pseudo) metric of distributions and maximizing the difference between the positive distribution and the negative one based on the metric. Accordingly, we extend (1) and reformulated as

$$\max\nolimits_f \mathbb{E}_{x\sim p_{\mathcal{X}}}[d(p_{\mathcal{P}|x},p_{\mathcal{N}|x};f)],\tag{3}$$

where $d(p,q;f)$ is the metric defined for the distributions associated with the model f. From this viewpoint, we can interpret (1) with more possibilities. For example, when $h*$ is an identity function, $i.e.$, $h^*(t)=t$, and (1) becomes a score matching framework and the metric d is the maximum mean discrepancy (MMD) (Dziugaite, Roy, and Ghahramani 2015; Li et al. 2017).

Spherical sliced Wasserstein distance

In this work, we specialize the contrastive learning framework in (3) based on the theory of optimal transport (Villani 2008). Given two distributions p and q defined on the sample space $\mathcal{X}$, the Wasserstein distance between them is defined as

$$\begin{aligned}d_{\mathrm{w}}(p,q):=&\min_{\pi\in\Pi_{p,q}}\int_{\mathcal{X}\times\mathcal{X}}d_{\mathcal{X}}(x,x')\pi(x,x')dxdx'\\ &=\min\nolimits_{\pi\in\Pi_{p,q}}\mathbb{E}_{x,x'\sim\pi}[d_{\mathcal{X}}(x,x')],\end{aligned}\tag{4}$$

where $\Pi_{p,q}$ is the set of the joint distributions using p and q as marginals, $d_{\mathcal{X}}$ is the metric of the sample space $\mathcal{X}$.

A naïve way to implement the contrastive learning framework in (3) is using $d_w(p_{\mathcal{P}|x},p_{\mathcal{N}|x};f)$ directly, $i.e.$, $\min_{\pi_{p_{\mathcal{P}|x},p_{\mathcal{N}|x}}}\mathbb{E}_\pi[|f(x_j)-f(x_{j'})|^2]$. However, two problems need to be solved: (i) how to leverage the conditional information provided by the sample x; and (ii) how to make the Wasserstein distance itself more computationally-friendly.

To solve these two problems, we proposed a spherical sliced Wasserstein (SSW) distance for the contrastive learning problem. In practice, the Wasserstein distance is often replaced with an equivalent surrogate called sliced Wasserstein distance (Bonneel et al. 2015; Kolouri, Rohde, and Hoffmann 2018):

$$d_{\mathrm{sw}}(p,q):=\mathbb{E}_{u\in\mathcal{S}^{M-1}}[d_{\mathrm{w}}(p_{\#u},q_{\#u})],\tag{5}$$

where u is a random projection sampled from the M-dimensional sphere $\mathcal{S}^{M-1}$, $p_{\#u}$ is the 1D distribution of the samples projected along the direction u, and

$$d_{\mathrm{w}}(p_{\#u},q_{\#u}):=\min\nolimits_{\pi\in\Pi_{p_{\#u},q_{\#u}}}\mathbb{E}_\pi[|u^\top x-u^\top x'|^2].\tag{6}$$

Our spherical sliced Wasserstein distance further parametrized the distribution of the random projection. Instead of sampling the projection u uniformly from $\mathcal{S}^{M-1}$, we sample u from a power spherical distribution (Nguyen et al. 2020):

$$\begin{aligned} & u \sim p(u; z), \\ & p(u; z) = \frac{1}{2^{M-1+\kappa}\pi^{\frac{d-1}{2}}\frac{\Gamma(\frac{M-1}{2}+\kappa)}{\Gamma(M-1+\kappa)}}(1 + z^\top u)^\kappa, \end{aligned} \tag{7}$$

where $\kappa \geq 0$ is the concentration parameter, $z \in \mathcal{S}^{M-1}$ is the location vector. Accordingly, our spherical sliced Wasserstein distance is defined as

$$d_{\text{ssw}}(p, q) := \max\nolimits_{z\in\mathcal{S}^{M-1}} \mathbb{E}_{u\sim p(u;z)}[d_{\text{w}}(p_{\#u}, q_{\#u})]. \tag{8}$$

Plugging (8) into (3), we obtain a new paradigm of contrastive learning:

$$\max\nolimits_{f,g} \mathbb{E}_{x\sim p_{\mathcal{X}}} \mathbb{E}_{u\sim p(u;g(x))}[d_{\text{w}}(p_{\mathcal{P}\#u|x}, p_{\mathcal{N}\#u|x}; f)] \tag{9}$$

Given a batch of positive and negative samples corresponding to the sample x, denoted as $\mathcal{P}_x$ and $\mathcal{N}_x$, we can implement the objective function as

$$\sum_{n,k=1}^{N,K} \min_{T\in\Pi_{1,1}} \sum_{i,j=1}^{|\mathcal{P}_{x_n}|,|\mathcal{N}_{x_n}|} |u_k^\top f(x_i) - u_k^\top f(x_j)|^2 t_{ij}, \tag{10}$$

where N is the number of samples, $u_k \sim p(u; g(x_n))$ is the k-th projection sampled from $p(u; g(x_n))$. Here, $g : \mathcal{X} \mapsto \mathcal{S}^{M-1}$ maps the conditional sample x_n to $\mathcal{S}^{M-1}$, which determines the direction of the projection.

When $|\mathcal{P}_x| = |\mathcal{N}_x|$, we can further rewrite the objective function as

$$\sum_{n,k=1}^{N,K} \sum_{i,j=1}^{|\mathcal{P}_{x_n}|,|\mathcal{N}_{x_n}|} |\text{sort}(u_k^\top f(x_i)) - \text{sort}(u_k^\top f(x_j))|^2. \tag{11}$$

Learning Algorithm

Alternating optimization strategy

Reparametrization of projector

Related Work

Contrastive learning

Wasserstein distance

Experiments

Image representation

MNIST, CelebA

Backbone model can be ResNet or some other well-known models (pls check existing contrastive learning work.)

Node embedding and clustering

References

Bonneel, N.; Rabin, J.; Peyré, G.; and Pfister, H. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51(1): 22–45.

Dziugaite, G. K.; Roy, D. M.; and Ghahramani, Z. 2015. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, 258–267.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2018. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Kolouri, S.; Rohde, G. K.; and Hoffmann, H. 2018. Sliced wasserstein distance for learning gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3427–3436.

Li, C.-L.; Chang, W.-C.; Cheng, Y.; Yang, Y.; and Póczos, B. 2017. MMD GAN: towards deeper understanding of moment matching network. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2200–2210.

Nguyen, K.; Nguyen, S.; Ho, N.; Pham, T.; and Bui, H. 2020. Improving Relational Regularized Autoencoders with Spherical Sliced Fused Gromov Wasserstein. In *International Conference on Learning Representations*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, 271–279.

Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2018. Deep Graph Infomax. In *International Conference on Learning Representations*.

Villani, C. 2008. *Optimal transport: old and new*, volume 338. Springer Science & Business Media.