An Optimal Transport Approach to Deep Metric Learning (Student Abstract)

Jason Xiaotian Dou[1], **Lei Luo**[1*], **Raymond Mingrui Yang**[2]

[1] Department of Electrical and Computer Engineering, University of Pittsburgh
[2] Department of Electrical and Computer Engineering, Carnegie Mellon University
jason.dou@pitt.edu, leiluo2017@gmail.com, mingruiy@andrew.cmu.edu

Abstract

Capturing visual similarity among images is the core of many computer vision and pattern recognition tasks. This problem can be formulated in such a paradigm called metric learning. Most research in the area has been mainly focusing on improving the loss functions and similarity measures. However, due to the ignoring of geometric structure, existing methods often lead to sub-optimal results. Thus, several recent research methods took advantage of Wasserstein distance between batches of samples to characterize the spacial geometry. Although these approaches can achieve enhanced performance, the aggregation over batches definitely hinders Wasserstein distance's superior measure capability and leads to high computational complexity. To address this limitation, we propose a novel Deep Wasserstein Metric Learning framework, which employs Wasserstein distance to precisely capture the relationship among various images under ranking-based loss functions such as contrastive loss and triplet loss. Our method directly computes the distance between images, considering the geometry at a finer granularity than batch level. Furthermore, we introduce a new efficient algorithm using Sinkhorn approximation and Wasserstein measure coreset. The experimental results demonstrate the improvements of our framework over various baselines in different applications and benchmark datasets.

Introduction

Learning a distance metric has been a key step for many applications in machine learning (Meng et al. 2019; Chen et al. 2019). In this work, we propose a novel Deep Wasserstein Metric Learning Framework that features the ability to learn beyond Euclidean space. The framework takes advantage of two new ranking based loss functions: wtriplet loss and wcontrastive loss, which are formulated to capture image similarities via Wasserstein distance. We further utilize the Sinkhorn approximation and Wasserstein measure coreset to reduce the high computational complexity of Wasserstein losses.

Method

We show a simple and elegant design is the best suit for Wasserstein loss in deep metric learning. We replace the Euclidean distance with Wasserstein distance in original contrastive loss and triplet loss by defining $d_w(x, y) = W_1(x, y)$. So the new Wasserstein-contrastive (wcontrastive) loss and Wasserstein-triplet (wtriplet) loss can be formulated as the following:

$$\begin{aligned}\mathcal{L}_{\text{wcontrastive}} = \frac{1}{b} \sum_{(i,j)\in\mathcal{P}}^{b} & I_{y_i=y_j} d_w\left(\phi_i, \phi_j\right) + \\ & I_{y_i\neq y_j}\left[\gamma - d_w\left(\phi_i, \phi_j\right)\right]_+\end{aligned} \tag{1}$$

$$\mathcal{L}_{\text{wtriplet}} = \frac{1}{b} \sum_{\substack{(a,p,n)\in\mathcal{T} \\ y_a=y_p\neq y_n}}^{b} \left[d_w\left(\phi_a, \phi_p\right) - d_w\left(\phi_a, \phi_n\right) + \gamma\right]_+ \tag{2}$$

In practice, direct computation of Wasserstein distance is way beyond the capability of our computational resources. We address this problem and improve efficiency by using the idea of Sinkhorn. The resulting entropic regularized p-Wasserstein distance is:

$$\begin{aligned}W^p_{p,\eta}(\mu, \nu) = \underset{\pi\in\Pi(\mu,\nu)}{\arg\min} \int_{\mathcal{X}\times\mathcal{X}} & \|x-y\|^p \mathrm{d}\pi(x, y) + \\ & \eta \mathrm{KL}(\pi \| \mu \otimes \nu)\end{aligned} \tag{3}$$

Regularizing the Wasserstein distance with an entropic penalty opens the door for new numerical approaches to compute optimal transport. (Genevay, Peyré, and Cuturi 2018) further introduces Sinkhorn divergences, defined as follows:

$$SD_{p,\eta}(\mu, \nu) = W_{p,\eta}(\mu, \nu) - \frac{1}{2}\left(W_{p,\eta}(\mu, \mu) + W_{p,\eta}(\nu, \nu)\right) \tag{4}$$

We use the Sinkhorn divergence as an approximation of Wasserstein distance in practice. However, experiments demonstrate Wasserstein losses take much more time comparing to the rest of loss. So we further take advantage of the coreset idea to reduce computational cost.

Most existing methods focus on coreset optimization with Euclidean space (Mirzasoleiman, Cao, and Leskovec 2020). Furthermore, this problem formation ignores the fact that a dataset is an empirical sample of data distribution, which describes a learning task. To build the bridge between coreset and Wasserstein distance, we introduce the idea of Wasserstein coreset (Claici, Genevay, and Solomon 2020) and follow the notations and definitions of coreset and measure coreset. Then we have the following proposition:

[*] Corresponding Author

Algorithm 1: WCRAIG (Wasserstein Coresets for Accelerating Incremental Gradient Descent)

```
**Input:** measure $\mu, n > 0$, minibatch size $m, \gamma > 0$
**Output:** Subset $S \subseteq V$
initialize $(x_1, \ldots, x_n) \sim \mu$
$S_0 \leftarrow \emptyset, s_0 = 0, i = 0$
$F_w(S) < L(\{s_0\}) - \epsilon \; j \in \arg\max_{e \in V \setminus S_{i-1}} F_w(e \mid S_{i-1})$
$S_i = S_{i-1} \cup \{j\}$
$i = i + 1$
Sample $(y_1, \ldots, y_m) \sim \mu$
Update estimate of $v^*$ using samples $y_k$.
Define generalized Voronoi regions $V_i(v^*)$.
Step: $x_i \leftarrow x_i - \gamma \nabla_x SD_{p,\eta}$
```

Proposition 1 (**Wasserstein Coresets**). *A sufficient condition for ν to be a ε-coreset for μ and $\mathcal{F}$ is $W_1(\mu, \nu) \leq \varepsilon$. Thus, a strategy for constructing an n-point coreset for a measure μ is to solve for $\{\mathbf{x}_i\}_{i=1}^n$*

$$\underset{\mathbf{x}_1, \ldots, \mathbf{x}_n}{\arg\min} W_1\left(\mu, \frac{1}{n}\sum_{i=1}^{n} \delta_{\mathbf{x}_i}\right) \tag{5}$$

Then we make an intuitive modification to the original CRAIG algorithm. in the submodular facility location function, we define

$$F_w(S) = \sum_{i \in V} \max_{j \in S} \frac{1}{SD_{p,n}(i, j)} \tag{6}$$

where we use the reciprocal of the Sinkhorn divergence to represent the similarity between two images $i, j \in V$ The procedure for WCRAIG is outlined in Algorithm 1.

Experiments

The experiments aim to demonstrate that our method can achieve superior performance compared to several representative deep metric learning methods on image retrieval and clustering tasks. We follow the experiment set up from (Roth et al. 2020). We summarize the ten loss functions' performance and highlight the best performance for each benchmark in Table 1. Overall, wtriplet achieves superior performance in most metrics against all the baseline loss functions, including the state-of-the-art loss functions: margin loss and multisimilarity loss (Roth et al. 2020).

Conclusion

We propose a novel Deep Wasserstein Metric Learning approach from optimal transport perspective, which offers new insight into deep metric learning. We take advantage of the Sinkhorn approximation and Wasserstein measure coreset to address the computational challenge. This framework offers several appealing benefits: it suggests a way to extend deep metric learning beyond the Euclidean setting. The method achieves competitive results on standard image retrieval and clustering benchmarks. The experimental results demonstrate the superiority of our framework over existing methods.

CUB200-2011 (Wah et al. 2011)					
Approach	R@1	R@4	NMI	F1	mAP
Wtriplet	**0.7277**	**0.8763**	**0.6876**	0.3610	**0.2474**
Wcontrastive	0.6094	0.8168	0.6720	**0.3615**	0.2395
Triplet	0.5949	0.7098	0.8023	0.2968	0.2357
Contrastive	0.5738	0.7919	0.6053	0.2906	0.1925
Npair	0.6241	0.8295	0.6676	0.3599	0.2332
ProxyNCA	0.6280	0.8190	0.6693	0.3610	0.2394
GenLifted	0.5959	0.7950	0.6563	0.3486	0.2203
Histogram	0.6055	0.8056	0.6526	0.3388	0.2265
Margin	0.6493	0.8489	0.6836	0.3593	0.2411
Multisimilarity	0.6280	0.8501	0.6855	0.3603	0.2258

Stanford Online Products (Song et al. 2016)					
Approach	R@1	R@4	NMI	F1	mAP
Wtriplet	**0.7400**	**0.8313**	**0.9023**	**0.3875**	**0.3995**
Wcontrastive	0.6989	0.8297	0.8911	0.3473	0.3733
Triplet	0.6094	0.8168	0.7720	0.3615	0.3695
Contrastive	0.6204	0.7918	0.7553	0.3237	0.3307
Npair	0.5976	0.8028	0.6325	0.3050	0.3166
ProxyNCA	0.6214	0.8242	0.8761	0.3220	0.3058
GenLifted	0.7321	0.8193	0.8984	0.3593	0.3903
Histogram	0.7130	0.8205	0.8993	0.3158	0.3488
Margin	0.7352	0.8289	0.8903	0.3836	0.3836
Multisimilarity	0.7399	0.8301	0.8900	0.3675	0.3852

Table 1: Information Retrieval and Cluster Performance

References

Chen, S.; Luo, L.; Yang, J.; Gong, C.; Li, J.; and Huang, H. 2019. Curvilinear distance metric learning. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 4223–4232.

Claici, S.; Genevay, A.; and Solomon, J. 2020. Wasserstein Measure Coresets. arXiv:1805.07412.

Genevay, A.; Peyré, G.; and Cuturi, M. 2018. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, 1608–1617.

Meng, Y.; Huang, J.; Wang, G.; Zhang, C.; Zhuang, H.; Kaplan, L.; and Han, J. 2019. Spherical text embedding. *Advances in Neural Information Processing Systems*, 32: 8208–8217.

Mirzasoleiman, B.; Cao, K.; and Leskovec, J. 2020. Coresets for Robust Training of Deep Neural Networks against Noisy Labels. *Advances in Neural Information Processing Systems*, 33.

Roth, K.; Milbich, T.; Sinha, S.; Gupta, P.; Ommer, B.; and Cohen, J. P. 2020. Revisiting training strategies and generalization performance in deep metric learning. In *International Conference on Machine Learning*, 8242–8252. PMLR.

Song, H. O.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *Computer Vision and Pattern Recognition (CVPR)*.

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. *California Institute of Technology*, (CNS-TR-2011-001).