IEICE TRANS. INF. & SYST., VOL.E107–D, NO.7 JULY 2024

LETTER

Improving Sliced Wasserstein Distance with Geometric Median for Knowledge Distillation

Hongyun LU[†a)], **Mengmeng ZHANG**[††b)], **Hongyuan JING**[††], ***Nonmembers***, ***and*** **Zhi LIU**[†], ***Member***

SUMMARY Currently, the most advanced knowledge distillation models use a metric learning approach based on probability distributions. However, the correlation between supervised probability distributions is typically geometric and implicit, causing inefficiency and an inability to capture structural feature representations among different tasks. To overcome this problem, we propose a knowledge distillation loss using the robust sliced Wasserstein distance with geometric median (GMSW) to estimate the differences between the teacher and student representations. Due to the intuitive geometric properties of GMSW, the student model can effectively learn to align its produced hidden states from the teacher model, thereby establishing a robust correlation among implicit features. In experiment, our method outperforms state-of-the-art models in both high-resource and low-resource settings.

key words: *sliced Wasserstein, geometric median, knowledge distillation*

1. Introduction

In recent years, due to the rapid development of artificial intelligence, model compression has received a great deal of attention from researchers, especially regarding deep neural networks [1]. Knowledge distillation is one of the most commonly used methods in model compression, which is to transfer the dark knowledge in the complex model to the simple model. Hinton [2] introduced the idea of knowledge distillation in neural networks, which involves using a teacher model's high-level features as supervision for a smaller, more efficient student model. The teacher model is highly capable, while the student model is designed to achieve comparable prediction results with less complexity by metric learning, ideally approaching or even surpassing the teacher's performance. In the context of knowledge distillation, the metric learning is comprised of a linear combination of two distinct losses: the cross-entropy (CE) loss with "hard" targets, and the distillation loss with divergence of "soft" distributions. In multi-task learning scenarios involving local features such as image segmentation and object detection, the predicted probability distribution captures more informative and intricate geometric features. As a result, the Kullback-Leibler divergence and L2 loss as knowledge distillation loss [3] may not effectively convey significant geometric information in such contexts.

Manuscript received November 24, 2023.
Manuscript publicized March 8, 2024.
[†]The authors are with North China University of Technology, Beijing, China.
[††]The authors are with Beijing Union University, Beijing, China.
a) E-mail: herrylu@126.com (Corresponding author)
b) E-mail: muchmeng@126.com (Corresponding author)
DOI: 10.1587/transinf.2023EDL8083

The robust sliced Wasserstein distance [4], [5] with geometric median [6] projection (GMSW) for knowledge distillation is proposed in this paper, a new knowledge distillation loss that reduces the generalization gap between teacher and student to approach better knowledge transfer. Firstly, the sliced Wasserstein distance possesses a geometric interpretation, rendering it a suitable metric for comparing distributions with structural properties. Secondly, the sliced Wasserstein distance with geometric median exhibits superior resistance to outliers and noise compared to alternative distance metrics, including the Euclidean distance and KL divergence. The GMSW maps geometric feature in highly capable distribution of the teacher model to the student model through robust geometric median projections, which improves the performance of transfer learning. The method is validated on multiple models, achieving good results.

The remainder of the paper is organized as follows. A brief description of knowledge distillation and Wasserstein distance in Sect. 2. In Sect. 3, we describe the proposed method. The performance of the proposed method is presented in Sect. 4. Finally, Sect. 5 provides the conclusion.

2. Knowledge Distillation and Wasserstein Distance

2.1 Knowledge Distillation

Knowledge distillation is to transfer the dark knowledge in the complex model to the simple model, a student network is trained by leveraging additional supervision from a trained teacher network. Given an input sample (x, y), where x is the network input and y is the one-hot label.

$$P_t = softmax(Z_t(x)),\ P_s = softmax(Z_s(x)) \tag{1}$$

Assume Z_t and Z_s are the logit representations (before the SoftMax layer) of the teacher and student network and P is the output distribution in Eq. (1), respectively. The distillation objective encourages the output probability distribution over predictions from the student and teacher networks to be similar by minimizing the cross-entropy loss and knowledge distillation loss between predictions of teacher and student as follows:

$$Loss = H(P_s, y) + \lambda KD(P_s^\tau, P_t^\tau) \tag{2}$$

Where τ is a relaxation hyperparameter (referred as Temperature) for softening the output of teacher network, and λ is a

hyper-parameter for balancing cross-entropy and knowledge distillation loss Eq. (2). The idea of knowledge distillation is to let the student mimic the teacher's behavior by adding a strong congruent constraint on predictions using knowledge distillation loss.

2.2 Wasserstein Distance

In this section, we review the initial concepts and optimality conditions for computing the p-Wasserstein distance between two discrete probability measures in Monge and Kantorovich formulation for Wasserstein distance [7]. Let $P(R^d)$ be the set of Borel probability measures in R^d, and let $P_2(R^d)$ be the subset of $P(R^d)$ consisting of probability measures that have finite second moments. The p-Wasserstein distance in Eq. (3), $p \in [1, \infty)$, between μ and ν is defined as the solution of the optimal mass transportation problem. Let μ and ν be two probability measures on measurable spaces R^d. $\mu,\ \nu \in P_2(R^d)$ and $\Pi(\mu, \nu)$ be the set of couplings between μ, ν. For $\mu,\ \nu \in P_2(R^d)$, we write $\Pi(\mu, \nu)$, that satisfies the following.

$$\Pi(\mu,\nu) = \begin{cases} \pi(A \times R_d) = \mu(A) & \text{any Borel } A \subseteq R_d \\ \pi(R_d \times B) = \nu(B) & \text{any Borel } B \subseteq R_d \end{cases} \tag{3}$$

In model compression applications one often deals with compact d-dimensional Euclidean spaces, hence X = Y = $[0,1]^d$. The p-Wasserstein distance int Eq. (4) for $p \in [1, \infty)$ is defined as

$$W(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} (x-y)^p d\pi(x,y) \right)^{\frac{1}{p}} \tag{4}$$

Because we only consider low level costs in the rest of this paper, we will only use W_2 to denote the 2-Wasserstein distance. For $X,\ Y \subseteq R^d$ and T: $X \to$ Y, the push-forward of $\mu \in P$(X) by T is defined by $T\#\mu \to p(Y)$. In other words, $T\#\mu$ is the measure satisfying $T\#\mu(A) = \mu(T^{-1}(A))$ for any Borel set in Y. The 2-Wasserstein distance in Eq. (5) is defined by

$$W_2(\mu,\nu) := \left(\inf_{\pi \in \Pi(\mu,\nu)} \int ||x-y||^2 d\pi(x,y) \right)^{\frac{1}{2}} \tag{5}$$

3. The GMSW Loss

The sliced Wasserstein distance with geometric median projection is designed as knowledge distillation loss. The sliced Wasserstein distance is calculated via linear slicing of the probability distributions. In order to ensure an efficient evaluation of the sliced Wasserstein distance, a more informative projection is extracted by selectively linearizing these projections through geometric median. The geometric median is a statistical measure of central tendency that is determined by calculating the point that minimizes the sum of distances to all other points. It is also referred to as the geometric center or the median of a geometric distribution. The geometric median is less influenced by outliers or extreme values than the arithmetic or other means. Figure 1 pictorially illustrates our overall knowledge distillation loss in the sliced Wasserstein distance with geometric median.

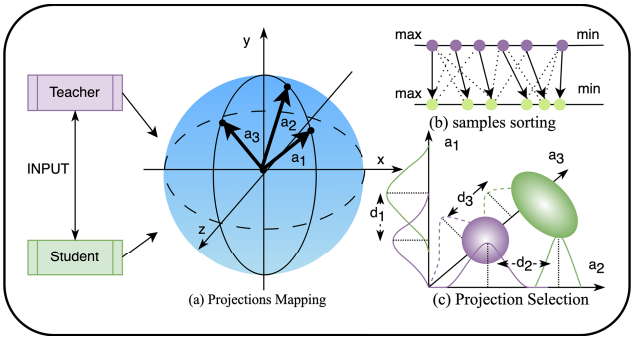

Fig. 1 The overview of the proposed GMSW knowledge distillation. Two backbone features of the teacher (purple) and the student (green). After inference, obtained features are projected onto the spherical plane in **a**. Samples in each projection are sorted and the distance is calculated between the sorted samples in **b**. The best projection distance from $\{d_1, d_2, d_3, \ldots\}$ as knowledge distillation loss is selected by the Geometric Median in **c**.

3.1 Sliced Wasserstein Distance for Knowledge Distillation

The sliced Wasserstein distance maps a high-dimensional probability distribution into a one-dimensional representation through projections, and then calculates the distance between two prediction distributions of teacher model and students' model as a functional on the p-Wasserstein distance of their one-dimensional presentation. The p-Wasserstein distance has a closed-form solution for the case of one-dimensional continuous probability measures. The slice process is related to the field of Integral Geometry and specifically the Radon transform. The relevant result to our discussion is that a d-dimensional probability density can be uniquely represented as the set of its one-dimensional marginal distributions following the Radon transform and the Fourier slice theorem. $\delta(\mu)$ denotes the one dimensional Dirac delta function, and $(\cdot, \cdot)$ denotes the Euclidean inner-product.

Definition 1 For any $\mu,\ \nu \in P_2(R^d)$, the SW distance of order 1 between them is defined as

$$SW_1(\mu,\nu) := \int_{S^{d-1}} W_1(\mu^*_{\#}\mu, \mu^*_{\#}\nu) d\delta(\mu) \tag{6}$$

For any $\mu \in S_{\mathrm{d}-1}$, let μ^* be the linear form with respect to u under the projection on θ, such that for $\theta \in R^d$, $\mu^*(\theta) = <\mu, \theta>$, δ represents the uniform distribution on $S_{\mathrm{d}-1}$. In the knowledge distillation application, the sliced Wasserstein distance need to be used for discrete measures. Since the expectation in Definition 1 is intractable, the Monte Carlo estimation is used projecting directions of length L.

$$\begin{cases} \mu^*_l(\theta) = (\theta_1, \theta_2, \theta_l, \ldots, \theta_L) \\ \widehat{SW}_1(\mu,\nu) \approx \frac{1}{L} \sum_{l=1}^{L} W_1(\mu^*_l\mu, \mu^*_l\nu) \end{cases} \tag{7}$$

$W_1(\mu_1^*\mu, \mu_1^*\nu)$ in Eq. (7) indicates that the empirical Wasserstein distance between $\mu_1^*\mu$ and $\mu_1^*\nu$ can be simply calculated by first sorting both samples and then culating the distance between the sorted samples.

3.2 Sliced Wasserstein Loss with Geometric Median

In this section, we discuss our approach that uses the Sliced Wasserstein Distance with Geometric Median (GMSW) to train a knowledge distillation model. In knowledge distillation, the role of distillation loss is to aggregate information about soft features, and its main contribution is to obtain important sources of geometric knowledge projection. we can capture the major discrepancy between two measures by considering a relatively small number of "important" slices. This problem can be alleviated by using Geometric Median, which considers as most representative projection in L projection directions. In the GMSW, the calculation can be simplified to pick the "best direction" along the projected distance which is the geometric median instead of using mean of the random projection directions generated from d dimensions for knowledge distillation. Based on this, the Sliced Wasserstein Distance with Geometric Median can be used as a more robust distance metric by replacing the Sliced Wasserstein Distance in its calculation.

Definition 2 Given a set of n positive real numbers $\{x_1, x_2, x_i, \ldots, x_n\}$, the geometric median defined as

$$\text{Geometric Median} = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\mathbf{x} - \mathbf{x_i}||_2 \tag{8}$$

Here, argmin means the value of the argument x which minimizes the sum. In this case, it is the point x in n dimensional Euclidean space from where the sum of all Euclidean distances to the x_i' is minimum. In GMSW, the distribution X in the Wasserstein distance with geometric median can be as the union vectors of all projected distances. the geometric median of all slices is the projection distance that minimizes the sum of its Euclidean distances to the other projection distance. More formally, following the notations in Eq. (9), the GMSW distance of order 1 is defined as

$$\widehat{GMSW}(\mu, \nu) \approx \underset{\mu^* \in \mathbb{R}^d}{\operatorname{argmin}} = \sum_{l=1}^{L} W_1(\mu_l^*\mu, \mu_l^*\nu) \tag{9}$$

The typical approach for computing geometric median is the Weiszfeld algorithm. The Weiszfeld algorithm is an iterative algorithm used to solve the geometric median problem. Its fundamental idea is to iteratively calculate the distance of each point to the mean point in order to approximate the geometric median value. The pseudo code of GMSW Loss is shown in Algorithm 1.

Algorithm 1 Sliced Wasserstein with Geometric Median

1: Input:$\{x_i \sim \mu\}_{i=1}^n$, $\{x_i \sim \nu\}_{i=1}^n$, slices number is L
2: Initialize:$D \longleftarrow 0$
3: for l = 1: L do
4: (i) Genetate a random vector μ_l from $\mathbb{S}^{d-1}$
5: (ii) Compute $\hat{x} = \langle \mu_l, x_i \rangle$ and $\hat{y} = \langle \mu_l, y_i \rangle$ for i=1,2,…,n
6: (iii) Sort $\{\hat{x}_i\}_{i=1}^n$and $\{\hat{y}_i\}_{i=1}^n$ denote by $\{\widehat{x_{|l|}}\}_{i=1}^n$ and $\{\widehat{y_{|l|}}\}_{i=1}^n$
7: (iv) $D_l = \{\widehat{x_{|l|}}\}_{i=1}^n - \{\widehat{y_{|l|}}\}_{i=1}^n$ and D += D_l
8: Initialize $m = {}^{D}/_{L}$ and a stopping criterion ε
9: for i = 1: Max_iterations do and for l = 1: L do
10: (i) Compute the distance $d_l = \|D_l - m\|$
11: (ii) Compute the weighted average $w_l = {}^{1}/_{d_l}$
12: (iii) Update numerator = $\sum(D_l * w_l)$
13: (iv) Update denominator = $\sum w_l$
14: (v) Update m_i=numerator/ denominator
15: (vi) Update deta = $\|m_i - m\|$
16: (vii) if deta<ε Return m as the geometric median

4. Experiments Results

We performed experiments on large-scale datasets: namely the CityScapes dataset [8]. We select three types of targets with thousands of images per class, which are pedestrians, vehicles, and motorbikes, and use the bounding rectangle of instance annotation as the bounding box of the target. In experiments, the performance of the GMSW loss with KL divergence and L2 loss are evaluated comparatively in object detection.

Table 1 Experimental results of our proposed method.

Type	Network	Loss	mAP
Normal	ResNet50	CE	27.6
Normal	ResNet18	CE	24.5
Normal	MobileNet	CE	21.2
Distillation	ResNet18	CE+KL divergence	25.4
Distillation	ResNet18	CE+L2	24.1
Distillation	**ResNet18**	**CE+GMSW**	**27.2**
Distillation	MobileNet	CE+KL divergence	22.9
Distillation	MobileNet	CE+L2	22.4
Distillation	**MobileNet**	**CE+GMSW**	**24.1**

Common Settings. The backbone network for teacher model in all experiments is ResNet-50. For the student structure, we use more compact ResNet18 amd MobileNet as well as its variants with different FLOPs, since ResNet18 and MobileNet have been proved to be highly effective in keeping high accuracy while maintaining low FLOPs in many tasks. We conduct all the experiments on a computer with 1 NVIDIA V100 GPUs, and the object detection framework is based on CenterNet in our experiments. The same number of slices ($L = 100$) is set for all comparable scores in GMSW loss. In Table 1, the term "Type" denotes the training approach, where "Normal" indicates the model achieved through standard training, "Distillation" refers to the student model trained with ResNet50 as the teacher model, and "Loss" specifically denotes the method of supervision in knowledge distillation. In our study, we adopt the evaluation metric of Average Precision, which is commonly referred to as mAP.

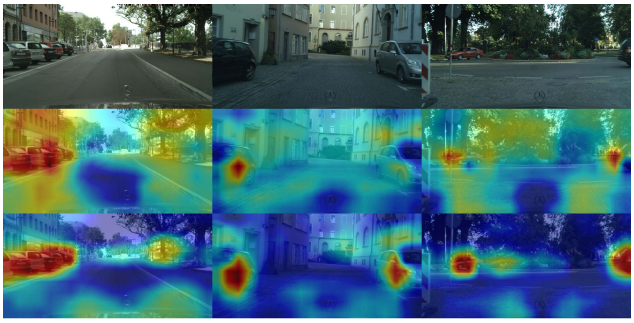

Fig. 2 Visualization of output features. The second row and the third row show the feature come from KL loss and GMSW loss, respectively. The heatmaps are highlighted, proving that GMSW loss distillation can make the detector focus on the geometric feature.

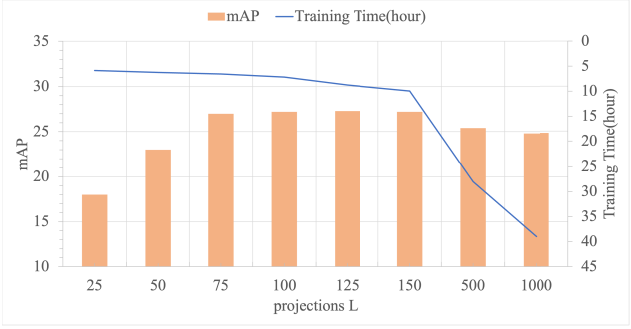

Fig. 3 Robust estimation and training time for the projections L.

Our results show that the GMSW Loss outperformed the L2 and KL divergence in terms of accuracy. Specifically, the GMSW loss achieved an average mAP of 27.2%, which significantly surpassed the L2 loss's 3.1% and the KL loss's 1.8% for the student model of ResNet18. Figure 2 shows the visualization results obtained from the student model of ResNet18. Our observations indicate that the geometric attributes of the foreground features in third row is more align closely with the object regions in the first row than KL loss. Furthermore, the GMSW mechanism facilitates the detector to emphasize the geometric properties of the interested objects, such as angle, scale, and size.

To achieve the better robustness performance, we analyzed the impact of the number of projections by adjusting L. As shown in Fig. 3, increasing the number of projections L did not significantly improve mAP and led to training instability and longer training time. This is due to the fact that, in the context of object detection distillation, the foreground typically constitutes a relatively small proportion in comparison to the background. It is necessary to limit the number of projections to avoid noise interference and improve robustness.

5. Conclusions

We study the focus on key learning aspect of the Geometric Median Sliced Wasserstein (GMSW) loss for knowledge distillation in this paper. Our work provides an enhanced understanding of sliced Wasserstein distances with geometric median and the associated minimal distance estimators under knowledge distillation. Our results suggest that GMSW loss can significantly improve the robustness and accuracy for knowledge distillation. Further research is needed to investigate the applicability of SW to high dimensional data and to explore the optimal parameters for SW.

Acknowledgments

This work is supported by MOE Planned Project of Humanities and Social Sciences (No.20YJA870014) and Beijing Natural Science Foundation (L223022).

References

[1] J. Gou, B. Yu, S.J. Maybank, and D. Tao, "Knowledge distillation: A survey," International Journal of Computer Vision, vol.129, pp.1789–1819, 2021.
[2] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," arXiv preprint arXiv:1503.02531, 2015.
[3] T. Kim, J. Oh, N. Kim, S. Cho, and S.-Y. Yun, "Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation," arXiv preprint arXiv:2105.08919, 2021.
[4] I. Deshpande, Y.-T. Hu, R. Sun, A. Pyrros, N. Siddiqui, S. Koyejo, Z. Zhao, D. Forsyth, and A.G. Schwing, "Max-sliced Wasserstein distance and its use for gans," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.10648–10656, 2019.
[5] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced Wasserstein distances," Advances in Neural Information Processing Systems, 2019.
[6] Y. He, P. Liu, Z. Wang, Z. Hu, and Y. Yang, "Filter pruning via geometric median for deep convolutional neural networks acceleration," Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp.4340–4349, 2019.
[7] C. Villani, Optimal transport: Old and new, Springer, Berlin, 2009.
[8] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," CVPR Workshop on the Future of Datasets in Vision, vol.2, 2015.