SUPPLEMENTARY MATERIAL FOR *"GENERALIZABLE EMBEDDINGS WITH CROSS-BATCH METRIC LEARNING"*

Yeti Z. Gürbüz[†] *A. Aydın Alatan*[‡]

[†]RSiM, Technische Universität Berlin, DE [‡]Center for Image Analysis (OGAM), METU, TR

Appendix

Preliminaries

Definition 1 (Optimal Transport Distance) *The optimal transport (OT) distance between two probability mass distributions (p, X) and (q, Y) is:*

$$\|(p, X) - (q, Y)\|_{OT} = \min_{\substack{\pi \geqslant 0 \\ \Sigma_i \pi_{ij} = q_j \\ \Sigma_j \pi_{ij} = p_i}} \sum_{ij} c_{ij} \pi_{ij} \tag{A.1}$$

*where $c_{ij} = \|x_i - y_j\|_2$, and $(p, X) \in \Sigma_n \times \mathbb{R}^{d\times n}$ denotes a probability mass distribution with masses $p \in \Sigma_n$ in the probability simplex (*i.e., $\Sigma_n \coloneqq \{p \in \mathbb{R}^n_{\geqslant 0} \mid \sum_i p_i = 1\}$*), and d-dimensional support $X = [x_i]_{i\in[n]} \in \mathbb{R}^{d\mathsf{x}n}$.*

Definition 2 (Maximum Mean Discrepancy) *Maximum mean discrepancy (MMD) between two probability mass distributions (p, X) and (q, Y) is:*

$$\|(p, X) - (q, Y)\|_{MMD} = \max_{f\in\mathcal{C}(X,Y)} \sum_i p_i f(x_i) - \sum_j q_j f(y_j) \tag{A.2}$$

where $\mathcal{C}(X, Y)$ is the set of continuous and bounded functions defined on a set covering the column vectors of X and Y.

Definition 3 (Optimal Transport Distance Dual) *The Lagrangian dual of the optimal transport distance defined in Definition 1 reads:*

$$\|(p, X) - (q, Y)\|_{OT} = \max_{f_i + g_j \leqslant c_{ij}} \sum_i p_i f_i + \sum_j q_j g_j \tag{A.3}$$

with the dual variables $\lambda = \{f, g\}$.

Note that $x_i = y_j$ implies $f_i = -g_j$ and from the fact that $c_{ij} = c_{ji}$, we can express the problem in (A.3) as:

$$\|(p, X) - (q, Y)\|_{OT} = \max_{f\in\mathfrak{L}_1} \sum_i p_i f(x_i) - \sum_j q_j f(x_j) \tag{A.4}$$

where $\mathfrak{L}_1 = \{f \mid \sup\limits_{x,y} \frac{|f(x)-f(y)|}{\|x-y\|_2} \leqslant 1\}$ is the set of 1-Lipschitz functions.

[†]Affiliated with OGAM-METU during the research.

Proofs

Definition 4 (Histogram Operator) *For n-many d-dimensional features $X=[x_i\in\mathbb{R}^d]_{i=1}^n$ and m-many prototype features $\mathcal{V} = [\nu_i\in\mathbb{R}^d]_{i=1}^m$ of the same dimension, the histogram of X on $\mathcal{V}$ is denoted as z^* which is computed as the minimizer of the following problem:*

$$(z^*, \pi^*) = \underset{z\in\mathcal{S}^m, \pi\geqslant 0}{\arg\max} \sum_{ij} \nu_i^\intercal x_j \pi_{ij} \;\; s.to \;\; \begin{matrix} \Sigma_i \pi_{ij} = 1/n \\ \Sigma_j \pi_{ij} = z_i \end{matrix} \tag{A.5}$$

where $\mathcal{S}^m \coloneqq \{p \in \mathbb{R}^m_{\geqslant 0} \mid \Sigma_i p_i = 1\}$.

Claim 1 *The solution of the problem in* (A.5) *reads:*

$$\pi^*_{ij} = 1/n \mathbb{1}(i = \mathrm{argmax}_k\{\nu_k^\intercal x_j\}) \tag{A.6}$$

where $\mathbb{1}(c)$ is 1 whenever c is true and 0 otherwise.

Proof: We prove our claim by contradiction. Denoting $c_{ij} = -\nu_i^\intercal x_j$, for any j, we express a solution as $\pi^*_{ij} = \epsilon_i$ with $\epsilon_i \geqslant 0$ and $\sum_i \epsilon_i = 1/n$. Let $i^* = \arg\min_k\{c_{kj}\}$. We can write $\pi^*_{i^*j} = 1/n - \sum_{i|i\neq i^*} \epsilon_i$. Our claim states that $\epsilon_i = 0$ for $i \neq i^*$. We assume an optimal solution, π', with $\epsilon_i > 0$ for some $i \neq i^*$. Since π' is optimal, we must have $\sum_{ij} \pi'_{ij} c_{ij} \leqslant \sum_{ij} \pi_{ij} c_{ij}$ for any π. For the j^{th} column we have,

$$\begin{aligned} \sum_i \pi'_{ij} c_{ij} &= \Big(\frac{1}{n} - \sum_{i'|i'\neq i^*} \epsilon_{i'}\Big) c_{i^*j} + \sum_{i'|i'\neq i^*} \epsilon_{i'} c_{i'j} \\ &= \frac{1}{n} c_{i^*j} + \sum_{i'|i'\neq i^*} \epsilon_{i'} (c_{i'j} - c_{i^*j}) \overset{(a)}{>} \sum_i \pi^*_{ij} c_{ij} \end{aligned}$$

where in (a) we use the fact that $(c_{i'j} - c_{i^*j}) > 0$ and $\epsilon_{i'} > 0$ for some i' by the assumption. Hence, $\sum_{ij} \pi'_{ij} c_{ij} > \sum_{ij} \pi^*_{ij} c_{ij}$ poses a contradiction. Therefore, $\epsilon_{i'} = 0$ must hold for all $i' \neq i^*$. ∎

Lemma 1 *Given n-many convolutional features $X = [x_i\in\mathcal{X}]_{i=1}^n$, and m-many prototype features $\mathcal{V} = [\nu_i]_{i=1}^m$ with $\{\nu_i\}_{i=1}^m$ being δ-cover of $\mathcal{X}$. If z^* is the histogram of X on V, defined in* (A.5)*, then we have:*

$$\|\sum_{i=1}^m z_i^* \nu_i - \sum_{j=1}^n \frac{1}{n} x_j\|_2 \leqslant \delta$$

Proof: We can express

$$\|\sum_{i\in[m]} z_i^*\nu_i - \sum_{j\in[n]} \frac{1}{n}x_j\|_2^2 = \sum_{i\in[m]} p_i^* f(\nu_i) - \sum_{j\in[n]} q_j f(x_j)$$

where $f(x) = x^\intercal(\sum_i z_i^*\nu_i - \sum_j \frac{1}{n}x_j)$, and $[n] = 1, \ldots, n$. Note that f is a continuous bounded operator for $\mathcal{X} = \{x \mid \|x\|_2 \leqslant 1\}$ (We can always map the features inside unit sphere without loosing the relative distances). Moreover, the operator norm of f, *i.e.* $\|f\|$, which is $\|\sum_i z_i^*\nu_i - \sum_j \frac{1}{n}x_j\|_2$ is less than or equal to 1. Thus, f lie in the unit sphere of the continuous bounded functions set. Using the definition of MMD distance, we can bound the error as:

$$\sum_{i\in[m]} z_i^* f(\nu_i) - \sum_{j\in[n]} q_j f(x_j) \leqslant \|(z^*, V) - (q, X)\|_{MMD}$$

where $q_i = 1/n$ for all i. For the continuous and bounded functions of the operator norm less than 1, MMD is lower bound for OT [1]. Namely,

$$\begin{aligned}\sum_{i\in[m]} z_i^* f(\nu_i) - \sum_{j\in[n]} q_j f(x_j) &\leqslant \|(z^*, V) - (q, X)\|_{MMD}\\ &\leqslant \|(z^*, V) - (q, X)\|_{OT}\end{aligned}$$

Since columns of V is δ-cover of the set $\mathcal{X}$, the optimal transport distance between the two distributions are bounded by δ, *i.e.* $\|(z^*, V) - (q, X)\|_{OT} \leqslant \delta$. Thus, we finally have:

$$\|\sum_{i\in[m]} z_i^*\nu_i - \sum_{j\in[n]} \frac{1}{n}x_j\|_2 \leqslant \delta.$$

■

1. REFERENCES

[1] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet, "Hilbert space embeddings and metrics on probability measures," *The Journal of Machine Learning Research*, vol. 11, pp. 1517–1561, 2010.