Supplementary material for "Similarity Learning with Top-heavy Ranking Loss for Person Re-identification"

Jin Wang, Nong Sang, Zheng Wang, Changxin Gao

I. Overview

The supplementary file consists of three sections. In Sec. I, the overview of this supplementary file is given. In Sec. II, we present the optimization details of our approach (SLTRL). In Sec. III, we present the optimization problem which is formulated with the triplet loss.

II. Details of the Optimization

After reformulating Eq.(6) into the unconstrained form of Eq.(11), the similarity learning problem can be solved by gradient descent algorithm. Due to the huge dimensionality of $\mathcal{Y}^{p,u}$, there are exponential number of constraints in Eq.(6). Joachims *et al*. [R1] showed that this problem can be efficiently solved by a cutting plane algorithm, *i.e*., only finds a small set of active constraints that ensures a sufficiently accurate solution. When using cutting plane approach to optimize the objective function, one key step is the separation oracle. Given fixed W and b, the separation oracle aims to find the most violated output $\tilde{\mathbf{y}}^{p,u}$:

$$\tilde{\mathbf{y}}^{p,u} \leftarrow \underset{\mathbf{y}^{p,u}\in\mathcal{Y}^{p,u}\setminus\hat{\mathbf{y}}^{p,u}}{\arg\max} \Delta_{TOP}(\mathbf{y}^{p,u},\hat{\mathbf{y}}^{p,u}) + \frac{1}{N}\cdot\sum_{v,k}(y^{p,u}_{v,k}-\hat{y}^{p,u}_{v,k})(s(\mathbf{x}^{p,u},\mathbf{x}^{g,v})-s(\mathbf{x}^{p,u},\mathbf{x}^{g,k})). \tag{S1}$$

As observed by Yue *et al*. [R2], optimizing over $\mathbf{y}^{p,u}$ is reduced to finding an optimal interleaving of the relevant and irrelevant sets. So (S1) could be efficiently calculated in $O(n\log n)$ time complexity, where $n=|\mathcal{R}^+|+|\mathcal{R}^-|$.

After selecting the most violated output $\tilde{\mathbf{y}}^{p,u}$, the loss caused by $\mathbf{x}^{p,u}$ and its most violated output $\tilde{\mathbf{y}}^{p,u}$ can be denoted as:

$$\ell(\mathbf{x}^{p,u},\tilde{\mathbf{y}}^{p,u})=[\Delta_{TOP}(\mathbf{y}^{p,u},\tilde{\mathbf{y}}^{p,u})+\frac{1}{N}\sum_{v,k}(\tilde{y}^{p,u}_{v,k}-y^{p,u}_{v,k})(f^T(\mathbf{x}^{p,u})f(\mathbf{x}^{g,v})-f^T(\mathbf{x}^{p,u})f(\mathbf{x}^{g,k}))]_+, \tag{S2}$$

where $[z]_+=\max(0,z)$ is the hinge loss function. And the total loss is rewritten as:

$$L=\frac{\lambda}{2}||\mathbf{W}||_F^2+\frac{1}{U}\sum_{u=1}^{U}\ell(\mathbf{x}^{p,u},\tilde{\mathbf{y}}^{p,u}). \tag{S3}$$

The gradient of L with respect to parameters $\mathbf{W}$ and $\mathbf{b}$ can be calculated as follows:

$$\frac{\partial L}{\partial \mathbf{W}} = \lambda \mathbf{W} + \frac{1}{U}\sum_{u=1}^{U}\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{W}} \tag{S4}$$

$$\frac{\partial L}{\partial \mathbf{b}} = \frac{1}{U}\sum_{u=1}^{U}\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{b}} \tag{S5}$$

Thus we can first calculate $\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{W}}$ (or $\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{b}}$), then add them together to get $\frac{\partial L}{\partial \mathbf{W}}$ (or $\frac{\partial L}{\partial \mathbf{b}}$). To make a clear illustration, we rewrite $f(\mathbf{x})$ in Eq.(2) as a composition of two functions:

$$f(\mathbf{x}) = g_2(g_1(\mathbf{x})), \tag{S6}$$

where $g_1(\mathbf{x}) = \tanh(\mathbf{W}^T\mathbf{x}+\mathbf{b})$ is the nonlinear transformation function and $g_2(\mathbf{x}) = \frac{\mathbf{x}}{||\mathbf{x}||_2}$ is the L2-normalization function. The gradient $\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{W}}$ can be calculated by the chain rule:

$$\begin{aligned}\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{W}} &= \frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial f(\mathbf{x}^{p,u})}\frac{\partial f(\mathbf{x}^{p,u})}{\partial g_1(\mathbf{x}^{p,u})}\frac{\partial g_1(\mathbf{x}^{p,u})}{\partial \mathbf{W}} \\ &+ \sum_{\mathbf{x}^{g,v}\in\mathcal{R}^+}\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial f(\mathbf{x}^{g,v})}\frac{\partial f(\mathbf{x}^{g,v})}{\partial g_1(\mathbf{x}^{g,v})}\frac{\partial g_1(\mathbf{x}^{g,v})}{\partial \mathbf{W}} \\ &+ \sum_{\mathbf{x}^{g,k}\in\mathcal{R}^-}\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial f(\mathbf{x}^{g,k})}\frac{\partial f(\mathbf{x}^{g,k})}{\partial g_1(\mathbf{x}^{g,k})}\frac{\partial g_1(\mathbf{x}^{g,k})}{\partial \mathbf{W}}\end{aligned} \tag{S7}$$

The gradient $\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{b}}$ can be calculated in a similar manner by replacing $\frac{\partial g_1(\cdot)}{\partial \mathbf{W}}$ with $\frac{\partial g_1(\cdot)}{\partial \mathbf{b}}$.

Gradients of the Ranking Loss Layer. Given a probe image $\mathbf{x}^{p,u}$, the gradients of $\ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})$ with respect to the image embedding inputs $f(\mathbf{x}^{p,u})$, $f(\mathbf{x}^{g,v})$, $f(\mathbf{x}^{g,k})$ are:

$$\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial f(\mathbf{x}^{p,u})} = \begin{cases} \frac{1}{N}\sum_{v,k}(\tilde{y}^{p,u}_{v,k} - \hat{y}^{p,u}_{v,k})(f(\mathbf{x}^{g,v}) - f(\mathbf{x}^{g,k})), & H(\tilde{\mathbf{y}}^{p,u}) > 0 \\ 0, & H(\tilde{\mathbf{y}}^{p,u}) \leq 0 \end{cases} \tag{S8}$$

For $\mathbf{x}^{g,v} \in \mathcal{R}^+(\mathbf{x}^{p,u})$,

$$\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial f(\mathbf{x}^{g,v})} = \begin{cases} \frac{1}{N}(\tilde{y}^{p,u}_{v,k} - \hat{y}^{p,u}_{v,k})f(\mathbf{x}^{p,u}), & H(\tilde{\mathbf{y}}^{p,u}) > 0 \\ 0, & H(\tilde{\mathbf{y}}^{p,u}) \leq 0 \end{cases} \tag{S9}$$

For $\mathbf{x}^{g,k} \in \mathcal{R}^-(\mathbf{x}^{p,u})$,

$$\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial f(\mathbf{x}^{g,k})} = \begin{cases} -\frac{1}{N}(\tilde{y}^{p,u}_{v,k} - \hat{y}^{p,u}_{v,k})f(\mathbf{x}^{p,u}), & H(\tilde{\mathbf{y}}^{p,u}) > 0 \\ 0, & H(\tilde{\mathbf{y}}^{p,u}) \leq 0 \end{cases} \tag{S10}$$

Gradients of the L2-normalization Layer. The gradient of the L2-normalization output with respect to the input is:

$$\frac{\partial f(\mathbf{x})}{\partial g_1(\mathbf{x})} = \frac{1}{||g_1(\mathbf{x})||_2}\mathbf{I} - \frac{1}{||g_1(\mathbf{x})||_2^{3/2}}g_1(\mathbf{x})g_1^T(\mathbf{x}) \tag{S11}$$

where $g_1(\mathbf{x}) \in \mathcal{R}^d$ and $\mathbf{I}$ is the identity matrix of size d. $g_1(\mathbf{x})$ is a single layer neural network, thus its gradient with respect to $\mathbf{x}$ can be easily calculated by the back propagation algorithm. The overall algorithm is presented in Alg.(1).

[R1] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, Large margin methods for structured and interdependent output variables, in Journal of Machine Learning Research, 2005, pp. 1453−1484.

[R2] Y. Yue, T. Finley, F. Radlinski, and T. Joachims, A support vector method for optimizing average precision, in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, 2007, pp. 271−278.

Algorithm 1: Learning similarity function with top-heavy ranking loss

Input: Training probe set: $\mathcal{P} = \{\mathbf{x}_u^p\}_{u=1}^U$, gallery set: $\mathcal{G} = \{\mathbf{x}_v^g\}_{v=1}^V$, learning rate η, iterative number T, regularization parameter λ

Output: The network parameters $\mathbf{W}, \mathbf{b}$

Initialize $\mathbf{W}, \mathbf{b}$;

for $t = 1, 2, ..., T$ **do**

// Forward propagation: calculate embedding features of each training image

for $u = 1, 2, ..., U$ **do**

Calculate $f(\mathbf{x}^{p,u})$ according to Eq.(S6);

end

for $v = 1, 2, ..., V$ **do**

Calculate $f(\mathbf{x}^{g,v})$ according to Eq.(S6);

end

// Back propagation: calculate gradients with respect to parameters

for $u = 1, 2, ..., U$ **do**

Solving (S1) to obtain the most violated output $\tilde{\mathbf{y}}^{p,u}$ for $\mathbf{x}^{p,u}$;

Calculate $\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{W}}$ and $\frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{b}}$ according to Eq.(S7)$\sim$ Eq.(S11);

Accumulate the gradient: $\frac{\partial L}{\partial \mathbf{W}} = \frac{\partial L}{\partial \mathbf{W}} + \frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{W}}$, $\frac{\partial L}{\partial \mathbf{b}} = \frac{\partial L}{\partial \mathbf{b}} + \frac{\partial \ell(\mathbf{x}^{p,u}, \tilde{\mathbf{y}}^{p,u})}{\partial \mathbf{b}}$;

end

Calculate $\frac{\partial L}{\partial \mathbf{W}}$ according to Eq.(S4);

Calculate $\frac{\partial L}{\partial \mathbf{b}}$ according to Eq.(S5);

Update network parameters: $\mathbf{W}^t = \mathbf{W}^{t-1} - \eta \frac{\partial L}{\partial \mathbf{W}}$, $\mathbf{b}^t = \mathbf{b}^{t-1} - \eta \frac{\partial L}{\partial \mathbf{b}}$;

end

Return $\mathbf{W}, \mathbf{b}$;

III. Optimization of Triplet Loss

For the triplet loss, we solve the following problem:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b}} \tfrac{1}{2}||\mathbf{W}||_F^2 + \tfrac{1}{U}\sum_{u=1}^{U} \xi_u \\ &s.t. \quad s(\mathbf{x}^{p,u}, \mathbf{x}^{g,v}) - s(\mathbf{x}^{p,u}, \mathbf{x}^{g,k}) > g - \xi_u, \xi_u \geq 0, \\ &\mathbf{x}^{g,v} \in \mathcal{R}^+(\mathbf{x}^{p,u}),\ \mathbf{x}^{g,k} \in \mathcal{R}^-(\mathbf{x}^{p,u}), \text{for } \forall \mathbf{x}^{p,u} \in \mathcal{P}, \end{aligned} \tag{S12}$$

where g is a margin parameter and is set to 0.5 in the experiment. The constraint in Eq.(S12) forces the similarity between relevant images to be larger than that of irrelevant images at least by a margin g. This problem is solved in a similar way with Eq.(6), *i.e.*, first reformulate it as an unconstrained form, then apply gradient descent algorithm to solve it.