Supplementary Material
Moving in the Right Direction: A Regularization for Deep Metric Learning

Deen Dayal Mohan*, Nishant Sankaran*, Dennis Fedorishin,
Srirangaraj Setlur, Venu Govindaraju,
Department of Computer Science and Engineering, University at Buffalo, Buffalo, New York, USA

1. Derivation of Direction Regularization

Following the motivation presented in Section 3.2 of the main manuscript where we state our objective of having the negative sample move in a direction orthogonal to both the anchor and positive samples, we require:

$$NC \perp PA \implies \frac{NC}{\|NC\|} \cdot \frac{PA}{\|PA\|} = 0 \tag{1}$$

Since we know that $NC = \boldsymbol{f}_c - \boldsymbol{f}_n$ and $PA = \boldsymbol{f}_a - \boldsymbol{f}_p$ from Fig. 3, we want

$$\frac{(\boldsymbol{f}_c - \boldsymbol{f}_n)}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|} \cdot \frac{(\boldsymbol{f}_a - \boldsymbol{f}_p)}{\|\boldsymbol{f}_a - \boldsymbol{f}_p\|} = 0 \tag{2}$$

$$\left[\frac{\boldsymbol{f}_c}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|} - \frac{\boldsymbol{f}_n}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|}\right] \cdot \left[\frac{\boldsymbol{f}_a - \boldsymbol{f}_p}{\|\boldsymbol{f}_a - \boldsymbol{f}_p\|}\right] = 0 \tag{3}$$

Expanding using the distributive laws of the inner product, we get:

$$\begin{aligned} &\frac{\boldsymbol{f}_c}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|} \cdot \left[\frac{\boldsymbol{f}_a - \boldsymbol{f}_p}{\|\boldsymbol{f}_a - \boldsymbol{f}_p\|}\right] - \\ &\frac{\boldsymbol{f}_n}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|} \cdot \left[\frac{\boldsymbol{f}_a - \boldsymbol{f}_p}{\|\boldsymbol{f}_a - \boldsymbol{f}_p\|}\right] = 0 \end{aligned} \tag{4}$$

Now, since $\boldsymbol{f}_c$ is the midpoint of the displacement vector $PA = \boldsymbol{f}_a - \boldsymbol{f}_p$, it will be perpendicular to PA and so, $\boldsymbol{f}_c \cdot (\boldsymbol{f}_a - \boldsymbol{f}_p) = 0$. This zeroes out the first half of the above equation, resulting in:

$$\frac{\boldsymbol{f}_n}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|} \cdot \left[\frac{\boldsymbol{f}_a - \boldsymbol{f}_p}{\|\boldsymbol{f}_a - \boldsymbol{f}_p\|}\right] = 0 \tag{5}$$

Since the dot product is homogeneous under scaling, we can rewrite the above equation as:

$$\begin{aligned} &\frac{1}{\|\boldsymbol{f}_c - \boldsymbol{f}_n\|} \frac{1}{\|\boldsymbol{f}_a - \boldsymbol{f}_p\|} (\boldsymbol{f}_n \cdot [\boldsymbol{f}_a - \boldsymbol{f}_p]) = 0 \\ &\implies \boldsymbol{f}_n \cdot \boldsymbol{f}_a - \boldsymbol{f}_n \cdot \boldsymbol{f}_p = 0 \end{aligned}$$

which finally gives us Eq. 7 in Section 3.2 of the main manuscript. Adding and subtracting $\boldsymbol{f}_a \cdot \boldsymbol{f}_a - \boldsymbol{f}_p \cdot \boldsymbol{f}_a$ (note that $\boldsymbol{f}_a \cdot \boldsymbol{f}_a = 1$), we get

$$(\boldsymbol{f}_n - \boldsymbol{f}_a) \cdot (\boldsymbol{f}_p - \boldsymbol{f}_a) = 1 - \boldsymbol{f}_p \cdot \boldsymbol{f}_a \tag{6}$$

Now, we know that

$$\begin{aligned} Cos(AN, AP) &= Cos(\boldsymbol{f}_n - \boldsymbol{f}_a, \boldsymbol{f}_p - \boldsymbol{f}_a) \\ &= \frac{(\boldsymbol{f}_n - \boldsymbol{f}_a)}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|} \cdot \frac{(\boldsymbol{f}_p - \boldsymbol{f}_a)}{\|\boldsymbol{f}_p - \boldsymbol{f}_a\|} \end{aligned} \tag{7}$$

Therefore, we arrive at:

$$Cos(AN, AP) = \frac{1 - \boldsymbol{f}_p \cdot \boldsymbol{f}_a}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\| \|\boldsymbol{f}_p - \boldsymbol{f}_a\|} \tag{8}$$

This equation represents the optimal condition we desire to achieve. Hence, we find that minimizing this equation helps achieve the effect we are seeking with respect to having the negative embedding be orthogonal to both anchor and positive samples.

2. Derivation of Gradient Dynamics for Triplet Pair

Considering a triplet of samples comprising only of the anchor, positive and negative, we would like to analyze the effect of the gradient on the positions of the samples in the embedding space. We start with the triplet loss formulation including the direction regularization:

$$\begin{aligned} \mathcal{L}_{apn} =& \|\boldsymbol{f}_a - \boldsymbol{f}_p\|^2 - \|\boldsymbol{f}_a - \boldsymbol{f}_n\|^2 + \alpha \\ &- \gamma\, Cos(\boldsymbol{f}_n - \boldsymbol{f}_a, \boldsymbol{f}_p - \boldsymbol{f}_a) \\ \implies \mathcal{L}_{apn} =& \|\boldsymbol{f}_a - \boldsymbol{f}_p\|^2 - \|\boldsymbol{f}_a - \boldsymbol{f}_n\|^2 + \alpha \\ &- \gamma \frac{1 - \boldsymbol{f}_p \cdot \boldsymbol{f}_a}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\| \|\boldsymbol{f}_p - \boldsymbol{f}_a\|} \end{aligned} \tag{9}$$

Computing gradients with respect to the anchor embedding $\boldsymbol{f}_a$, we get:

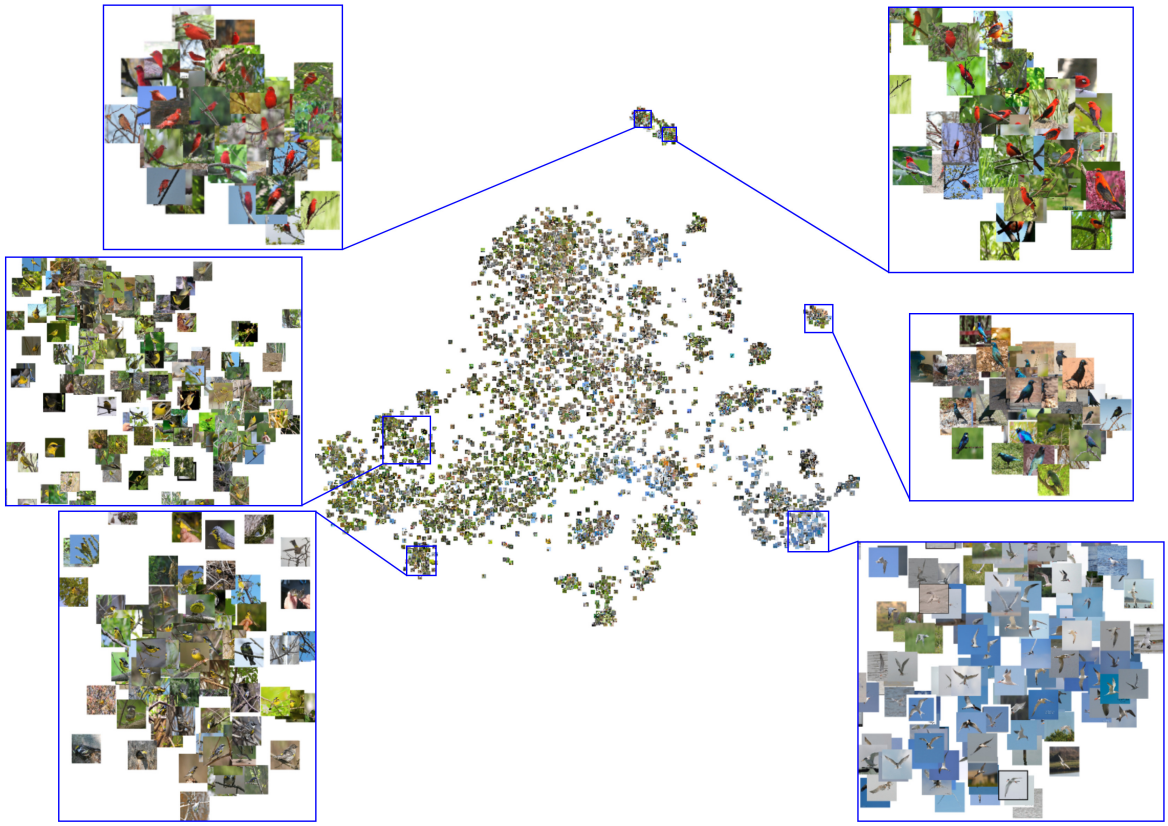

Figure 1: Visualization of DR-MS feature embedding of CUB-200-2001 (class 101 to 200; 5,924 images) using Barnes-Hut t-SNE. Various clusters are zoomed in for viewing and this figure is best viewed in color.

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_a} =& 2(\boldsymbol{f}_a - \boldsymbol{f}_p) - 2(\boldsymbol{f}_a - \boldsymbol{f}_n) - \frac{2\gamma(\boldsymbol{f}_a - \boldsymbol{f}_p)}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|\|\boldsymbol{f}_p - \boldsymbol{f}_a\|} \\ &+ \frac{\gamma}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|^2} \frac{\|\boldsymbol{f}_p - \boldsymbol{f}_a\|}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|}(\boldsymbol{f}_n - \boldsymbol{f}_a) \\ &+ \frac{\gamma(\boldsymbol{f}_a - \boldsymbol{f}_p)}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|\|\boldsymbol{f}_p - \boldsymbol{f}_a\|}\end{aligned} \tag{10}$$

Assigning $c = (\|\boldsymbol{f}_n - \boldsymbol{f}_a\|\|\boldsymbol{f}_a - \boldsymbol{f}_p\|)^{-1}$, $d = \|\boldsymbol{f}_n - \boldsymbol{f}_a\|^{-2}$ and $k = \|\boldsymbol{f}_n - \boldsymbol{f}_a\|^{-1}\|\boldsymbol{f}_a - \boldsymbol{f}_p\|$, we can simplify the gradient to:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_a} = 2(\boldsymbol{f}_n - \boldsymbol{f}_p) - \gamma c\,(\boldsymbol{f}_a - \boldsymbol{f}_p) - \gamma dk(\boldsymbol{f}_n - \boldsymbol{f}_a) \tag{11}$$

Similarly, we now compute the gradients with respect to the positive sample $\boldsymbol{f}_p$:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_a} =& -2(\boldsymbol{f}_a - \boldsymbol{f}_p) - \frac{2\gamma(\boldsymbol{f}_p - \boldsymbol{f}_a)}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|\|\boldsymbol{f}_p - \boldsymbol{f}_a\|} \\ &+ \frac{\gamma(\boldsymbol{f}_p - \boldsymbol{f}_a)}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|\|\boldsymbol{f}_p - \boldsymbol{f}_a\|}\end{aligned} \tag{12}$$

Using the same assignments for c, d, and k, we get

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_p} = 2(\boldsymbol{f}_p - \boldsymbol{f}_a) - \gamma c\,(\boldsymbol{f}_p - \boldsymbol{f}_a) \tag{13}$$

Computing the gradients with respect to the negative sample $\boldsymbol{f}_n$ yields:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_a} = 2(\boldsymbol{f}_a - \boldsymbol{f}_n) - \frac{\gamma}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|^2} \frac{\boldsymbol{f}_a - \boldsymbol{f}_n}{\|\boldsymbol{f}_n - \boldsymbol{f}_a\|\|\boldsymbol{f}_p - \boldsymbol{f}_a\|} \tag{14}$$

$$\implies \frac{\partial \mathcal{L}}{\partial \boldsymbol{f}_n} = 2(\boldsymbol{f}_a - \boldsymbol{f}_n) - \gamma c\,d\,(\boldsymbol{f}_a - \boldsymbol{f}_n) \tag{15}$$

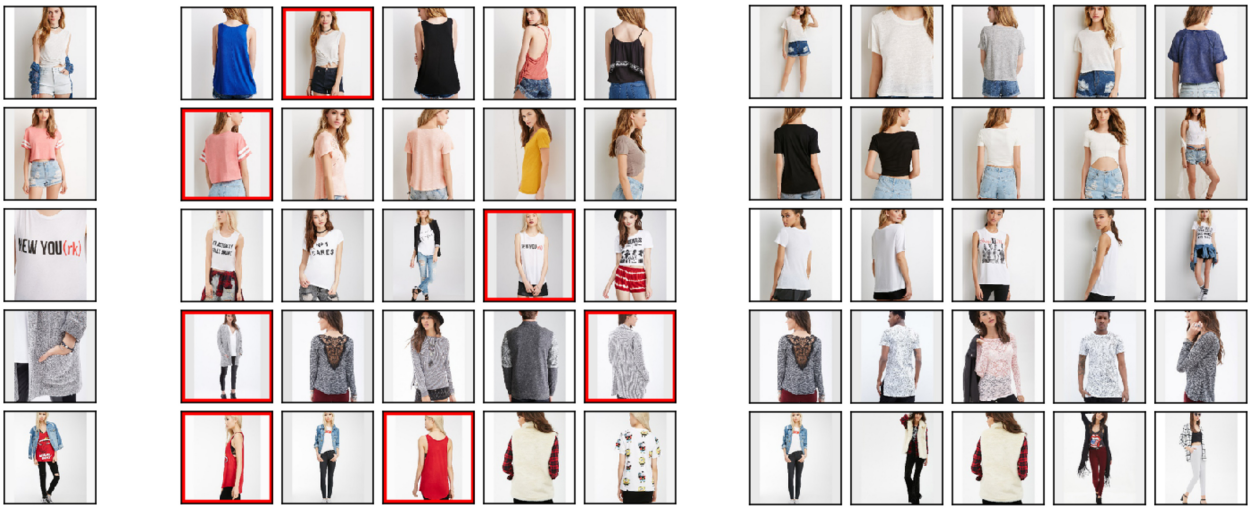

Figure 2: Qualitative results for top-5 recall on the In-Shop Clothes Retrieval dataset comparing the proposed DR-MS loss performance with MS-Loss. Images with a red border indicates the true positive gallery image for the given query image. As observed, DR-MS loss is able to correctly identify the mated gallery image for the requested query image in its top-5 candidate results as compared with MS loss.

Table 1: Study on the variation in performance on Caltech-UCSD CUB-200-2011 with respect to the embedding size employed.

Recall@K (%)	1
64	59.1
128	60.7
256	62.1
512	**66.1**

Which gives us the three gradients required to analyze their effects on the existing embedding positions. The remainder of the discussion follows in the main manuscript in Section 3.2

3. Embedding Size vs Performance

We analyse the performance variation of the metric learning system incorporated with direction regularization with respect to different embedding sizes. We use direction regularized MS-Loss for our experiment. We fix the batch size to 80 and use a learnable γ. The network and rest of hyper-parameters remains same as mentioned in the experiment section (§4). We report Recall@1 for the Caltech-UCSD CUB-200-2011. From Table 1 we note that the performance increases as the embedding dimension goes up. This is expected as higher dimensional features tend to have higher discriminative power. Importantly, we can observe that addition of the regularization term has not caused the performance variation behavior to have deviated from that of the original non-regularized formulation of the loss. This shows the versatility of the regularization term in how it can complement any metric learning loss without degrading its performance.