Rethinking preventing class-collapsing in metric learning with margin-based losses
Supplementary Material

Appendix

A: Proofs for the Theorems in subsection 4.2

We use all notions defined in subsection 4.2

Theorem 1. *Let $f : X \to \mathbb{R}^m$ be an embedding, which minimizes $\mathbb{EO}_{trip}(f)$, then f has the class-collapsing property with respect to all classes.*

Proof. Define a new random variables such that for every $1 \leq r_1, r_2 \leq t$:

$$h_{r_1,r_2}(Y,Z) = \begin{cases} 1 & Y = r_1 \wedge Z = r_2 \\ 0 & else \end{cases}$$

observe that

$$\bar{\delta}_{Y_1,Y_2} \cdot (1 - \bar{\delta}_{Y_1,Y_3}) = \sum_{\substack{1 \leq r_1, r_2 \leq t \\ r1 \neq r_2}} \mathbf{1}_{Y_1=r_1} \cdot h_{r_1,r_2}(Y_2, Y_3) = \sum_{\substack{1 \leq r_1, r_2 \leq t \\ r1 \neq r_2}} \mathbf{1}_{Y_1=r_2} \cdot h_{r_1,r_2}(Y_3, Y_2)$$

Since the variables are independent

$$\mathbb{E}(\bar{\delta}_{Y_1,Y_2} \cdot (1 - \bar{\delta}_{Y_1,Y_3})) = \frac{1}{2} \cdot \sum_{\substack{1 \leq r_1, r_2 \leq t \\ r1 \neq r_2}} \mathbb{E}(\mathbf{1}_{Y_1=r_1}) \cdot \mathbb{E}(h_{r_1,r_2}(Y_2, Y_3)) + \mathbb{E}(\mathbf{1}_{Y_1=r_2}) \cdot \mathbb{E}(h_{r_1,r_2}(Y_3, Y_2)).$$

Define: $\bar{D}(x_1, x_2, x_3) := (D_{x_1,x_2} - D_{x_1,x_3} + \alpha)_+$

Rearranging the terms we get

$$n^3 \cdot \mathbb{EO}_{trip}(f) = \sum_{x_1,x_2,x_3 \in X} (\mathbb{E}(\bar{\delta}_{Y_1,Y_2} \cdot (1 - \bar{\delta}_{Y_1,Y_3})) \cdot \bar{D}(x_1, x_2, x_3) =$$

$$\sum_{\substack{x_1,x_2,x_3 \in X \\ 1 \leq r_1 \neq r_2 \leq t}} (\mathbb{E}(\mathbf{1}_{Y_1=r_1}) \cdot \mathbb{E}(h_{r_1,r_2}((Y_2, Y_3)) + \mathbb{E}(\mathbf{1}_{Y_1=r_2}) \cdot \mathbb{E}(h_{r_1,r_2}(Y_3, Y_2))) \cdot \bar{D}(x_1, x_2, x_3) =$$

$$\sum_{\substack{x_1,x_2,x_3 \in X \\ 1 \leq r_1 \neq r_2 \leq t}} \mathbb{E}(h_{r_1,,r_2}(Y_2, Y_3)) \cdot \left(\mathbb{E}(\mathbf{1}_{Y_1=r_1}) \cdot \bar{D}(x_1, x_2, x_3) + \mathbb{E}(\mathbf{1}_{Y_1=r_2}) \cdot \bar{D}(x_1, x_3, x_2)\right) =$$

Therefore, if

$$K(i, j, k, r_1, r_2) = \left(\mathbb{E}(\mathbf{1}_{Y_1=r_1}) \cdot \bar{D}(x_i, x_j, x_k) + \mathbb{E}(\mathbf{1}_{Y_1=r_2}) \cdot \bar{D}(x_1, x_k, x_j)\right),$$

then $\mathbb{EO}_{trip}(f)$ can be written as

$$\mathbb{EO}_{trip}(f) = \frac{1}{n^3} \sum_{\substack{1 \leq i,j,k \leq n \\ 1 \leq r_1 \neq r_2 \leq t}} \mathbb{E}(h_{r_1,r_2}(Y_j, Y_k)) \cdot K(i, j, k, r_1, r_2)$$

For every $x_i \in X$, define:

$$(\mathbb{EO}_{trip}(f))_{x_i} = \frac{1}{n^2} \cdot \sum_{\substack{1 \le j,k \le n \\ 1 \le r_1 \ne r_2 \le t}} (\mathbb{E}(h_{r_1,r_2}(Y_j, Y_k)) \cdot K(x_i, x_j, x_k, r_1, r_2)$$

Let $f : X \to \mathbb{R}^m$ be an embedding, fix $1 \le r \le t$ and $x_i \in A_r, \quad x_j, x_k \in X$ with

$$\| f(x_i) - f(x_j)\| = w, \quad \| f(x_i) - f(x_k)\| = h.$$

By definition:

$$K(i,j,k,r_1,r_2) = \begin{cases} p \cdot (h - w + \alpha)_+ + q(w - h + \alpha)_+ & r_1 = r \wedge r_2 \ne r \\ q \cdot (h - w + \alpha)_+ + p(w - h + \alpha)_+ & r_2 = r \wedge r_1 \ne r \\ p \cdot (h - w + \alpha)_+ + p(w - h + \alpha)_+ & r_1 = r \wedge r_2 = r \\ q \cdot (h - w + \alpha)_+ + q(w - h + \alpha)_+ & r_1 \ne r \wedge r_2 \ne r \end{cases}$$

Since $0.5 < p < 1$, in order to get minimal $K(i,j,k,r_1,r_2)$ value, h and w must satisfy $|h - w| \le \alpha$. In this case we have

$$K(i,j,k,r_1,r_2) = \begin{cases} (p+q) \cdot \alpha + (h - w)(p - q) & r_1 = r \wedge r_2 \ne r \\ (p+q) \cdot \alpha + (w - h)(p - q) & r_2 = r \wedge r_1 \ne r \\ 2 \cdot \alpha & r_1 = r \wedge r_2 = r \\ 2 \cdot \alpha & r_1 \ne r \wedge r_2 \ne r \end{cases}$$

Therefore,

$$\sum_{r_2 \in \{1,..r-1,r+1,.t\}} (\mathbb{E}(h_{r,r_2}(Y_j, Y_k)) \cdot K(x_i, x_j, x_k, r_1, r_2) + (\mathbb{E}(h_{r_2,r}(Y_j, Y_k)) \cdot K(x_i, x_j, x_k, r_1, r_2) =$$
$$= (p+q) \cdot \alpha(\sum_{r_2 \in \{1,..r-1,r+1,.t\}} (\mathbb{E}(h_{r,r_2}(Y_j, Y_k) + (\mathbb{E}(h_{r_2,r}(Y_j, Y_k)))+$$
$$(h - w)(p - q))(\sum_{r_2 \in \{1,..r-1,r+1,.t\}} \mathbb{E}(h_{r,r_2}(Y_j, Y_k)) - \mathbb{E}(h_{r_2,r}(Y_j, Y_k))$$

We split to three cases:

1. If $x_j, x_k \in A_r$ or $x_j, x_k \notin A_r$ then: $\mathbb{E}(h_{r,r_2}(Y_j, Y_k)) = \mathbb{E}(h_{r_2,r}(Y_j, Y_k))$. Hence,

$$(h - w)(p - q))(\sum_{r_2 \in \{1,..r-1,r+1,.t\}} \mathbb{E}(h_{r,r_2}(Y_j, Y_k)) - \mathbb{E}(h_{r_2,r}(Y_j, Y_k)) = 0$$

2. If $x_j \in A_r$ and $x_k \notin A_r$, then $\mathbb{E}(h_{r,r_2}(Y_j, Y_k)) > \mathbb{E}(h_{r_2,r}(Y_j, Y_k))$, therefore

$$(h - w)(p - q))(\sum_{r_2 \in \{1,..r-1,r+1,.t\}} \mathbb{E}(h_{r,r_2}(Y_j, Y_k)) - \mathbb{E}(h_{r_2,r}(Y_j, Y_k))$$

   Since $p > 0.5$ and $|h - w| \le \alpha$,the minimal value is achieved whenever $h = 0$ and $w = \alpha$.

3. In the same way if $x_k \in A_r$ and $x_j \notin A_r$, then $\mathbb{E}(h_{h_{r_2,r}}(Y_j, Y_k)) = \mathbb{E}(h_{r,r_2}(Y_j, Y_k))$ and the minimal value is achieved whenever $h = \alpha$ and $w = 0$.

In conclusion, if $x_i \in A_r$, an embedding f^* satisfies

$$(\mathbb{EO}_{trip}(f^*))_{x_i} = \min\{(\mathbb{EO}_{trip}(f))_{x_i} | f : X \to \mathbb{R}^m\}$$

iff $f^*(x_j) = f^*(x_i)$ for every $x_j \in A_r$, and $\| f^*(x_j) - f^*(x_i)\| = \alpha$ for every $x_j \notin A_r$. □

We will now prove the same theorem with respect to the margin loss.

Theorem 2. *Let $f : X \to \mathbb{R}^m$ be an embedding, which minimizes*

$$\mathbb{EO}_{margin}(f,\beta) = \frac{1}{n^2}\sum_{x_i,x_j\in X}\mathbb{E}\mathcal{L}^f_{margin}(x_i,x_j),$$

then f has the class-collapsing property with respect to all classes.

Proof. Observe that if $x_i, x_j \in A_r$, then

$$\mathbb{E}\mathcal{L}^f_{margin}(x_i,x_j) = p\cdot(D_{x_i,x_j}-\beta_{x_i}+\alpha)_+ + (1-p)\cdot(\beta_{x_i}-D_{x_i,x_j}+\alpha)_+$$

Since $0 < p < 1$, then the maximal value is achieved whenever $|D_{x_i,x_j}-\beta_{x_i}| \le \alpha$, in this case:

$$\mathbb{E}\mathcal{L}^f_{margin}(x_i,x_j) = (2p-1)\cdot(D_{x_i,x_j}-\beta_{x_i}).$$

In the same way in case $x_i \in A_r$ and $x_j \notin A_r$ then:

$$\mathbb{E}\mathcal{L}^f_{margin}(x_i,x_j) = (2p-1)\cdot(\beta_{x_i}-D_{x_i,x_j}).$$

Combining both directions we get:

$$\sum_{x_j\in X}\mathbb{E}\mathcal{L}^f_{margin}(x_i,x_j) = (2p-1)\cdot\left(\sum_{Y_j\in A}D_{x_i,x_j}-\sum_{Y_j\notin A}D_{x_i,x_j}\right)$$

Since: $p > 0.5$ and $|D_{x_i,x_j}-\beta_{x_i}| \le \alpha$, the minimal value is achieved whenever $D_{x_i,x_j} = 0$, $D_{x_i,x_k} = 2\alpha$ and $\beta_{x_i} = \alpha$, for every $x_i, x_j \in A_r, \quad x_k \notin A_r$. □

B: Easy Positive Sampling in noisy environment

In this subsection we analyse the EPS method from the theoretical prospective, using the framework defined in Section 4. We use the same notions as in sections 3 and 4.

Define: $\Phi(y_i,y_j) = \begin{cases}1 & y_i = y_j \wedge D_{x_i,x_j} = min\{D_{x_i,x_k}|\, y_k = y_i\}\\ 0 & else\end{cases}$. Then, the easy positive sampling loss can be defined by:

$$\frac{1}{n}\sum_{1\le i,j,k\le n}\Phi(y_i,y_j)\cdot\mathcal{L}^f_{trip}(x_i,x_j,x_k)$$

for the triplet loss and

$$\frac{1}{n}\sum_{1\le i,j\le n}(\Phi(y_i,y_j)\cdot\mathcal{L}^{f,\beta}_{margin}(x_i,x_j)) + 1_{y_i\neq y_j}\mathcal{L}^{f,\beta}_{margin}(x_i,x_j)$$

for the margin loss.

In the noisy environment stochastic case, using section 4 notions, Φ becomes a random variable:

$$\bar{\Phi}(Y_i,Y_j) = \begin{cases}1 & Y_i = Y_j \wedge \forall t\,((D_{x_i,,x_t} < D_{x_i,x_j}) \to Y_t \neq Y_i)\\ 0 & else\end{cases}$$

Therefore, the triplet loss with EPS in the noisy environment case, becomes:

$$\mathbb{E}\mathcal{L}^f_{EPStrip}(x_i,x_j,x_k) = \mathbb{E}\left(\bar{\Phi}(Y_i,Y_j)\cdot\bar{\delta}_{Y_i,Y_j}\cdot(1-\bar{\delta}_{Y_i,Y_k})\right)\cdot\left(D^f_{x_i,x_j}-D^f_{x_i,x_k}+\alpha\right)_+$$

and for the margin loss with EPS we have:

$$\mathbb{E}\mathcal{L}^f_{EPSmargin}(x_i,x_j) = \mathbb{E}(\bar{\Phi}(Y_i,Y_j)\cdot\bar{\delta}_{Y_i,Y_j})\cdot(D^f_{x_i,x_j}-\beta_{x_i}+\alpha)_+ + \mathbb{E}(1-\bar{\delta}_{Y_i,Y_j})\cdot(\beta_{x_i}-D^f_{x_i,x_j}+\alpha)_+$$

As in section 4.2 we assume that $Y = \{Y_1,..,Y_n\}$ is a set of independent binary random variables. Let $A_1,..,A_t \subset X$, $0.5 < p < 1$ such that: $|A_j| = \frac{n}{t}$ and

$$\mathbb{P}(Y_i = k) = \begin{cases}p & x_i \in A_k\\ q' := \frac{1-p}{t-1} & x_i \notin A_k\end{cases}$$

For simplicity we assume that every $1 \leq i \leq \frac{n}{t}$ satisfies $x_{\frac{n \cdot i}{t}+1}, .., x_{\frac{n \cdot i}{t}+t} \in A_i$

We prove first that the minimal embedding with respect to both losses does not satisfy the class collapsing property. Let f_1 be an embedding function such that:

$$D^{f_1}_{x_i,x_j} = \begin{cases} 0 & (\exists r)(x_i, x_j \in A_r) \\ \alpha & else \end{cases}$$

and f_2 an embedding such that:

$$D^{f_1}_{x_1,x_2} = \begin{cases} 0 & (\exists r)(x_i, x_j \in A_r) \wedge \sim ((i < \frac{t}{2n} \wedge j > \frac{t}{2n}) \vee (i > \frac{t}{2n} \wedge j < \frac{t}{2n}) \\ \alpha & else \end{cases}$$

f_1 represent the case of class collapsing, where f_2 represent the case where there are two modalities for the first class. In order to show that the minimal embedding does not satisfy the class collapsing property it suffices to prove that

$$\frac{1}{n} \sum_{1 \leq i,j,k \leq n} \mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_i, x_j, x_k) < \frac{1}{n} \sum_{1 \leq i,j,k \leq n} \mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_i, x_j, x_k)$$

and

$$\frac{1}{n} \sum_{1 \leq i,j \leq n} \mathbb{E}\mathcal{L}^{f_2}_{EPSmargin}(x_i, x_j) < \frac{1}{n} \sum_{1 \leq i,j \leq n} \mathbb{E}\mathcal{L}^{f_1}_{EPSmargin}(x_i, x_j).$$

Remark: For both losses the definition requires a strict order between the elements, therefore by distance zero, we meant infinitesimal close, the order between the elements inside the sub-clusters is random, and element between set A_1 are closer then set A_1^c in both embeddings. For simplification we neglect this infinitesimal constants in the proofs.

Claim 1. *There exists M such that if $n \geq M$, then:*

$$\frac{1}{n} \sum_{1 \leq i,j,k \leq n} \mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_i, x_j, x_k) < \frac{1}{n} \sum_{1 \leq i,j,k \leq n} \mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_i, x_j, x_k)$$

Proof. Fix x_1, WOLOG we may assume in both embeddings that $D^{f_j}_{x_1,x_i} < D^{f_j}_{x_1,x_k}$ for every $j \in \{1, 2\}$ and $1 \leq i < k \leq n$. It suffices to prove that

$$\frac{1}{n} \sum_{1 \leq j,k \leq n} (\mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_1, x_j, x_k) - \mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_1, x_j, x_k)) > 0$$

Let $q = (1 - p)$, observe that

$$\mathbb{P}(\bigwedge_{1 \leq t < j} Y_i \neq Y_t) = p^{m+1} \cdot q^{j-2-m} + p^{j-2-m} q^{j+1} \leq 2p^{j-1}$$

where $m = |\{t \,|\, t \leq j,\ Y_t \in A_1\}|$. Thus if $j \geq \frac{n}{2t}$, we have

$$\mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_1, x_j, x_k) \leq \mathbb{P}(\bigwedge_{1 \leq t < j} Y_i \neq Y_t) \cdot 2 \cdot \alpha \leq 4 \cdot \alpha p^{j-1}.$$

Therefore,

$$\frac{1}{n} \sum_{j > \frac{n}{2t}, 1 \leq k \leq n} \mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_1, x_j, x_k) \leq \sum_{j > \frac{n}{2t}} 4 \cdot \alpha p^j = 4 \cdot n \cdot \alpha p^{\frac{n}{2t}} \cdot \sum_{j=0}^{\frac{n(2t-1)}{2t}} p^j = 4 \cdot \alpha p^{\frac{n}{2t}} \cdot \frac{1 - q^{n(2t-1)/2t}}{1 - q} \overset{n \to \infty}{\to} 0$$

For $j \leq \frac{n}{2t}$ and $k \leq \frac{n}{2t}$ of $k > \frac{n}{t}$, we have $\mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_1, x_j, x_k) = \mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_1, x_j, x_k)$. Hence, the only case left is $j \leq \frac{n}{2t}$ and $\frac{n}{2t} < k \leq \frac{n}{t}$. In this case: $\mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_1, x_j, x_k) = 0$, where

$$\mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_1, x_j, x_k) = (p^2 \cdot q^{j-1} + q^2 \cdot p^{j-1}) \cdot \alpha \geq q^{j+1}\alpha$$

and we get:

$$\frac{1}{n}\cdot\sum_{j\leq\frac{n}{2t},\frac{n}{2t}\leq k\leq\frac{n}{t}}\mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_1,x_j,x_k)-\mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_1,x_j,x_k)\geq$$

$$\alpha\cdot q^2\cdot\sum_{j=0}^{\frac{n}{2t}}q^i=\alpha\cdot q^2\cdot\frac{1-q^{n/2t}}{1-q}\overset{n\to\infty}{\to}\alpha q^2\cdot\frac{1}{1-q}$$

Choosing M such that

$$\alpha\cdot q^2\cdot\frac{1-q^{M/2t}}{1-q}>4\cdot\alpha p^{\frac{M}{4}}\cdot\frac{1-q^{M(2t-1)/2t}}{1-q}$$

will satisfy that for every $n>M$:

$$\frac{1}{n}\sum_{1\leq i,j,k\leq n}\mathbb{E}\mathcal{L}^{f_2}_{EPStrip}(x_i,x_j,x_k)<\frac{1}{n}\sum_{1\leq i,j,k\leq n}\mathbb{E}\mathcal{L}^{f_1}_{EPStrip}(x_i,x_j,x_k)$$

□

Claim 2. *There exists M such that if $n\geq M$ then:*

$$\frac{1}{n}\sum_{1\leq i,j\leq n}\mathbb{E}\mathcal{L}^{f_2}_{EPSmargin}(x_i,x_j)<\frac{1}{n}\sum_{1\leq i,j\leq n}\mathbb{E}\mathcal{L}^{f_1}_{EPSmargin}(x_i,x_j)$$

Proof. For every $1\leq j\leq\frac{n}{2t}$ or $\frac{n}{t}<j\leq n$ we have:

$$\mathbb{E}\mathcal{L}^{f_1}_{EPSmargin}(x_i,x_j)=\mathbb{E}\mathcal{L}^{f_1}_{EPSmargin}(x_i,x_j)$$

For $\frac{n}{2t}<j\leq\frac{n}{2}$:

$$\mathbb{E}\mathcal{L}^{f_2}_{EPSmargin}(x_i,x_j)=2\cdot p\cdot q\cdot\beta_{x_i}+(p^2q^{j-2}+q^2p^{j-2})\cdot(2\cdot\alpha-\beta_{x_i})$$

while:

$$\mathbb{E}\mathcal{L}^{f_2}_{EPSmargin}(x_i,x_j)=2\cdot p\cdot q\cdot(\beta_{x_i}+\alpha)+(p^2q^{j-2}+q^2p^{j-2})\cdot(\alpha-\beta_{x_i})$$

Since $j>\frac{n}{2t}$ the second therm tend to zero. Therefore, taking M such that

$$2qp>(p^2q^{\frac{M}{2t}-2}+q^2p^{\frac{M}{2t}-2})$$

will satisfy that for each $n\geq M$

$$\frac{1}{n}\sum_{1\leq i,j\leq n}\mathbb{E}\mathcal{L}^{f_2}_{EPSmargin}(x_i,x_j)<\frac{1}{n}\sum_{1\leq i,j\leq n}\mathbb{E}\mathcal{L}^{f_1}_{EPSmargin}(x_i,x_j)$$

□

In the previous two claims we prove that the class collapsing solution is not minimal with respect to both the $EPSmargin$ and the $EPStriplet$. In the following claims we prove that not only it is not the minimal solution, looking locally on the direct effect of the EPS losses on a sample which is not one of the closest elements to to the anchor. We prove that the optimal solution in this case is an embedding in which the distance between the sample to the anchor is equal to the margin hyperparameter.

Claim 3. *Let f be an embedding. For every i, let $i_1,..,i_n$ be such that $D^f_{x_i,x_{i_1}}<D^f_{x_i,x_{i_2}}<...<D^f_{x_i,x_n}$, Then there exists M such that for every $j>M$ the minimal embedding for $\mathbb{E}\mathcal{L}^f_{EPSmargin}(x_i,x_j)$ is achived whenever $D^f_{x_i,x_j}=\beta_{x_i}+\alpha$.*

Proof. Fix x_1, as in the previous claims we will assume:

$$D^f_{x_1,x_1} < D^f_{x_1,x_2} < ... < D^f_{x_1,x_n}$$

As was prove in in Claim 1 $\mathbb{P}(\bigwedge_{1\leq t<j} Y_i \neq Y_t) \leq 4p^j$, thus

$$\mathbb{E}(\bar{\Phi}(Y_i,Y_j)\cdot\bar{\delta}_{Y_i,Y_j}) \leq \mathbb{P}(\bigwedge_{1\leq t<j} Y_i \neq Y_t) \leq 4p^{j-1} \overset{j\to\infty}{\to} 0$$

Since the minimal solution for

$$\mathbb{E}(\bar{\Phi}(Y_i,Y_j)\cdot\bar{\delta}_{Y_i,Y_j})\cdot(D^f_{x_i,x_j}-\beta_{x_i}+\alpha)_+ + \mathbb{E}(1-\bar{\delta}_{Y_i,Y_j})\cdot(\beta_{x_i}-D^f_{x_i,x_j}+\alpha)_+$$

satisfies $|\beta_{x_i} - D^f_{x_i,x_j}| \leq \alpha$, we have:

$$\mathbb{E}\mathcal{L}^{f_1}_{EPSmargin}(x_1,x_j) = \alpha\cdot(\mathbb{E}(\bar{\Phi}(Y_1,Y_j)\cdot\bar{\delta}_{Y_1,Y_j}) + \mathbb{E}(1-\bar{\delta}_{Y_1,Y_j}))+$$
$$(D^f_{x_1,x_j}-\beta_{x_1})\cdot(\mathbb{E}(\bar{\Phi}(Y_1,Y_j)\cdot\bar{\delta}_{Y_1,Y_j}) - \mathbb{E}(1-\bar{\delta}_{Y_i,Y_j}))$$

Since $\mathbb{E}(1-\bar{\delta}_{Y_1,Y_j}) \geq 2pq$, there exists M such every $j > M$ satisfies

$$(\mathbb{E}(\bar{\Phi}(Y_1,Y_j)\cdot\bar{\delta}_{Y_1,Y_j}) - \mathbb{E}(1-\bar{\delta}_{Y_i,Y_j})) < 0$$

Therefore the minimal value is achieved whenever $D^f_{x_1,x_j} = \alpha + \beta_{x_1}$. □

The proof in the EPStriplet loss case is similar.

Claim 4. *Let f be an embedding. For every i, let $i_1,,..,i_n$ be such that $D^f_{x_i,x_{i_2}} < ... < D^f_{x_i,x_n}$. Then there exists M such that for every $j > M$ the minimal embedding for:*

$$\mathbb{E}\mathcal{L}^f_{EPStrip}(x_i,x_t,x_{t+j}) + \mathbb{E}\mathcal{L}^f_{EPStrip}(x_i,x_{t+j},x_t)$$

is achieved whenever $D_{x_i,x_{t+j}} = D_{x_i,x_t} + \alpha$.

Proof. Define $K(Y_i,Y_i,Y_k) := \mathbb{E}\left(\bar{\Phi}(Y_i,Y_j)\cdot\bar{\delta}_{Y_i,Y_j}\cdot(1-\bar{\delta}_{Y_i,Y_k})\right)$. Fixing x_1, assuming $D^f_{x_1,x_1} < D^f_{x_1,x_2} < ... < D^f_{x_1,x_n}$, We have:

$$\mathbb{E}\mathcal{L}^f_{EPStrip}(x_1,x_t,x_{t+j}) + \mathbb{E}\mathcal{L}^f_{EPStrip}(x_1,x_{t+j},x_t) = K(Y_1,Y_t,Y_{t+j})\cdot\left(D^f_{x_1,x_t} - D^f_{x_1,x_{t+j}} + \alpha\right)_+$$
$$+ K(Y_1,Y_{t+j},Y_t)\cdot\left(D^f_{x_1,x_{t+j}} - D^f_{x_1,t} + \alpha\right)_+$$

As in the previous claim, the minimal value is achieved whenever $|D^f_{x_1,x_{t+j}} - D^f_{x_1,x_t}| \leq \alpha$ in this case:

$$\mathbb{E}\mathcal{L}^f_{EPStrip}(x_1,x_t,x_{t+j}) + \mathbb{E}\mathcal{L}^f_{EPStrip}(x_1,x_{t+j},x_t) = \alpha\cdot(K(Y_1,Y_t,Y_{t+j}) + K(Y_1,Y_{t+j},Y_t))Y_t))+$$
$$(D^f_{x_1,x_t} - D^f_{x_1,x_{t+j}})\cdot(K(Y_1,Y_t,Y_{t+j}) - K(Y_1,Y_{t+j},Y_t))$$

On the one hand: $K(Y_1,Y_t,Y_{t+j}) = (\prod_{i\in\{1,2,..,t,t+j\}} p^{t_i}q^{1-t_i}) + (\prod_{i\in\{1,2,..,t,t+j\}} p^{1-t_i}q^{t_i}) \geq q^{t+1}$ where $t_k = \begin{cases}1 & Y_k \notin A\\ 0 & else\end{cases}$ for $k \in \{2,..,t-1,t+j\}$ and $t_k = \begin{cases}1 & Y_k \in A\\ 0 & else\end{cases}$ for $k \in \{1,t\}$.
On the other hand $K(Y_1Y_{t+j},Y_t) \geq Prob(\bigwedge_{1\leq k<t+j} Y_1 \neq Y_k) \leq 4p^{t+j-1}$. Taking j large enough such that $q^{t+1} \leq 4p^{t+j-1}$, we have:

$$(\mathbb{E}\left(\bar{\Phi}(Y_1,Y_t)\cdot\bar{\delta}_{Y_1,Y_t}\cdot(1-\bar{\delta}_{Y_1,Y_{t+j}})\right) - \mathbb{E}\left(\bar{\Phi}(Y_1,Y_{t+j})\cdot\bar{\delta}_{Y_1,Y_{t+j}}\cdot(1-\bar{\delta}_{Y_1,Y_t})\right)) > 0$$

therefore in such case the minimum is archived whenever $D^f_{x_1,x_{t+j}} = D^f_{x_1,x_t} + \alpha$.

□

dataset	model	Without EPS std	With EPS std
cars196	Margin	0.17	0.27
cars196	MS	0.24	0.29
cars196	Trip+SH	0.20	0.47
cub200	Margin	29.8	0.33
cub200	MS	0.43	0.36
cub200	Trip+SH	0.52	0.35
Omniglot-letters	Margin	0.73	0.58
Omniglot-letters	MS	0.52	0.71
Omniglot-letters	Trip+SH	0.34	0.61

Table 1: Std of Recall@1 results. Each model was trained 8 times with different random seeds.

	Cars196		CUB200	
	MS	MS+EPS	MS	MS+EPS
R@1	84.1	**85.5**	65.7	**66.7**
R@2	90.4	**90.7**	77.0	**77.2**
R@4	94.0	**94.3**	86.3	**86.4**
R@8	96.5	**96.7**	**91.2**	90.9

Table 2: Results of Multi-similarity loss with Embedding size 512 (as in [4]). Using EPS improve results in both cases.

C: More experiments and implementation details

MNIST architecture details

For the MNIST even/odd experiment we use a model consisting of two consecutive convolutions layer with (3,3) kernels and 32,64 (respectively) filter sizes. The two layers are followed by Relu activation and batch normalization layer, then there is a (2,2) max-pooling follows by 2 dense layers with 128 and 2 neurons respectively.

Stability analysis

Following [2, 1], it was important to us to have a fair comparison between all tested models. Therefore, for all the experiments we use the same framework as in [3], with the same architecture and embedding size (128). We also did not change the default hyper-parameters in all tested methods. We run each experiment 8 times with different random seeds, the reported results are the mean of all the experiments. The std of the Recall@1 results of all experiments can be seen in Table 1. In all cases the differences between the results with and without the EPS are significance.

Multi-similarity comparison

From our experiments, the Multi-similarity loss is highly affected by the batch size. Using Resnet50 backbone, we restrict the number of batch size to 160 for all tested model, which cause to the inferior results of the multi-similarity loss comparing to other methods. For the sake of completeness we provide the results also on inception backbone with embedding size of 512 as in [4], and batch size of 260. As can be seen in Table 2, also in these cases the results improve when using EPS instead of semi-hard sampling on the positive samples.

References

[1] Istvan Fehervari, Avinash Ravichandran, and Srikar Appalaraju. *Unbiased Evaluation of Deep Metric Learning Algorithms*. 2019. eprint: arXiv:1911.12528.

[2] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. *A Metric Learning Reality Check*. 2020. arXiv: 2003.08505 [cs.CV].

[3] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Björn Ommer, and Joseph Paul Cohen. “Revisiting Training Strategies and Generalization Performance in Deep Metric Learning”. In: (Feb. 19, 2020). arXiv: 2002.08473v7 [cs.CV].

[4] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. “Multi-Similarity Loss with General Pair Weighting for Deep Metric Learning”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019, pp. 5022–5030.