Simultaneous Similarity-based Self-Distillation for Deep Metric Learning

Karsten Roth\(^1\)\(^2\)  Timo Milbich\(^2\)  Björn Ommer\(^2\)  Joseph Paul Cohen\(^+\)\(^3\)  Marzyeh Ghassemi\(^+\)\(^4\)\(^1\)

Abstract

Deep Metric Learning (DML) provides a crucial tool for visual similarity and zero-shot applications by learning generalizing embedding spaces, although recent work in DML has shown strong performance saturation across training objectives. However, generalization capacity is known to scale with the embedding space dimensionality. Unfortunately, high dimensional embeddings also create higher retrieval cost for downstream applications. To remedy this, we propose Simultaneous Similarity-based Self-distillation (S2SD). S2SD extends DML with knowledge distillation from auxiliary, high-dimensional embedding and feature spaces to leverage complementary context during training while retaining test-time cost and with negligible changes to the training time. Experiments and ablations across different objectives and standard benchmarks show S2SD offers notable improvements of up to 7\% in Recall@1, while also setting a new state-of-the-art. Code available at https://github.com/MLforHealth/S2SD.

1. Introduction

Deep Metric Learning (DML) aims to learn embedding space models in which a predefined distance metric reflects not only the semantic similarities between training samples, but also transfers to unseen classes. The generalization capabilities of these models are important for applications in image retrieval (Wu et al., 2017), face recognition (Schroff et al., 2015), clustering (Bouchacourt et al., 2018) and representation learning (He et al., 2020). Still, transfer learning into unknown test distributions remains an open problem, with Roth et al. (2020b) and Musgrave et al. (2020) revealing strong performance saturation across DML training objectives. However, Roth et al. (2020b) also show that embedding space dimensionality can be a driver for generalization across objectives due to higher representation capacity. Indeed, this insight can be linked to recent work targeting other objective-independent improvements to DML via artificial samples (Zheng et al., 2019), higher feature distribution moments (Jacob et al., 2019) or orthogonal features (Milbich et al., 2020), which have shown promising relative improvements over selected DML objectives. Unfortunately, these methods come at a cost; be it longer training times or limited applicability. Similarly, drawbacks can be found when naively increasing the operating (base) dimensionality, incurring increased cost for data retrieval at test time, which is especially problematic on larger datasets. This limits realistically usable embedding dimensionalities and leads to benchmarks being evaluated against fixed, pre-defined dimensionalities.

In this work, we propose Simultaneous Similarity-based Self-Distillation (S2SD) to show that complex higher-dimensional information can actually be effectively leveraged in DML without changing the base dimensionality and test time cost, which we motivate from two key elements. Firstly, in DML, an additional embedding space can be spanned by a multilayer perceptron (MLP) operating over the feature representation shared with the base embedding space (see e.g. Milbich et al. (2020)). With larger dimensionalities, we can thus cheaply learn a secondary high-dimensional representation space simultaneously, also denoted as target embedding space. Relative to the large feature backbone, and with the batchsize capping the number of additional high dimensional operations, only little additional training cost is introduced. While we can not utilize the high-dimensional target embedding space at test time for aforementioned reasons, we may utilize it to boost the performance of the base embeddings. Unfortunately, a simple connection of base and additional target embedding spaces through the shared feature backbone is insufficient for the base representation space to benefit from the auxiliary, high-dimensional information. Thus, secondly, to efficiently leverage the high-dimensional context, we use insights from knowledge distillation (Hinton et al., 2015), where a small “student” model is trained to approximate a larger “teacher” model. However, while knowledge distillation can be found in DML (Chen et al., 2018), few-shot learning (Tian et al., 2020) and self-supervised extensions thereof (Rajasegaran et al., 2020), the reliance
on additional, commonly larger teacher networks or multiple training runs (Furlanello et al., 2018), introduces much higher training cost. Fortunately, we find that the target embedding space learned simultaneously at higher dimension can sufficiently serve as a “teacher” during training - through knowledge distillation of its sample similarities, the performance of the base embedding space can be improved notably. Such distillation intuitively encourages the lower-dimensional base embedding space to embed semantic similarities similar to the more expressive target embedding space and thus incorporate dimensionality-related generalization benefits.

Furthermore, S2SD makes use of the low cost to span additional spaces to introduce multiple teacher spaces. Operating each of them at higher, but varying dimensionality, joint distillation can then be used to enforce reusability in the distilled content akin to feature reusability in meta-learning (Raghu et al., 2020) for additional generalization boosts. Finally, in DML, the base embedding space is spanned over a penultimate feature space of much higher dimensionality, which introduces a dimensionality-based bottleneck (Milbich et al., 2020). By applying the distillation objective between feature and base embedding space in S2SD, we further encourage better feature usage in base embedding space. This facilitates the approximation of high-dimensional context through the base embedding space for additional improvements in generalization.

The benefits to generalization are highlighted in performance boosts across three standard benchmarks, CUB200-2011 (Wah et al., 2011), CARS196 (Krause et al., 2013) and Stanford Online Products (Oh Song et al., 2016), where S2SD improves test-set recall@1 of already strong DML objectives by up to 7%, while also setting a new state-of-the-art. Improvements are even more significant in very low dimensional base embedding spaces, making S2SD attractive for large-scale retrieval problems which can benefit from reduced embedding dimensionalities. Importantly, as S2SD is applied during the same DML training process on the same network backbone, no large teacher networks or additional training runs are required. Simple experiments even show that S2SD can outperform comparable 2-stage distillation at much lower cost.

In summary, our contributions can be described as:

1) We propose Simultaneous Similarity-based Self-Distillation (S2SD) for DML, using knowledge distillation of high-dimensional context without large additional teacher networks or training runs.
2) We motivate and evaluate this approach through detailed ablations and experiments, showing that the method is agnostic to choices in objectives, backbones, and datasets.
3) Across benchmarks, we achieve significant improvements over strong baseline objectives and state-of-the-art performance, with especially large boosts for very low-dimensional embedding spaces.

2. Related Work

Deep Metric Learning (DML) has proven useful for zero-shot image/video retrieval & clustering (Schroff et al., 2015; Wu et al., 2017; Brattoli et al., 2020), face verification (Liu et al., 2017; Deng et al., 2019) and contrastive (self-supervised) representation learning (e.g. He et al. (2020); Chen et al. (2020); Misra & van der Maaten (2020)). Approaches can be divided into 1) improved ranking losses, 2) tuple sampling methods and 3) extensions to the standard DML training approach. 1) Ranking losses place constraints on the relations available in image tuples, ranging from pairs (s.a. Hadsell et al. (2006)) to triplets (Schroff et al., 2015) and more complex orderings (Chen et al., 2017; Oh Song et al., 2016; Sohn, 2016; Wang et al., 2019). 2) As the number of possible tuples scales exponentially with dataset size, tuple sampling approaches have been introduced to tackle tuple redundancy and to ensure that meaningful tuples are presented during training. These tuple sampling methods can follow heuristics (Schroff et al., 2015; Wu et al., 2017), be of hierarchical nature (Ge, 2018) or learned (Roth et al., 2020a). Similarly, learnable proxies to replace tuple members (Movshovitz-Attias et al., 2017; Kim et al., 2020; Qian et al., 2019) can also remedy the sampling issue, which can be extended to tackle DML from a classification viewpoint (Zhai & Wu, 2018; Deng et al., 2019). 3) Finally, extensions to the basic training scheme can involve synthetic data (Lin et al., 2018; Zheng et al., 2019; Duan et al., 2018), complementary features (Roth et al., 2019; Milbich et al., 2020), a division into subspaces (Sanakoyeu et al., 2019; Xuan et al., 2018; Kim et al., 2018; Opitz et al., 2018), training of multiple networks (Park et al., 2020) using mutual learning (Zhang et al., 2018) or higher-order moments (Jacob et al., 2019). S2SD can similarly be seen as an extension to DML, though we specifically focus on capturing and distilling complex high-dimensional sample relations within lower dimensional embedding spaces to improve generalization.

Knowledge Distillation involves knowledge transfer from teacher to (usually smaller) student models, e.g. by matching network softmax outputs/logits (Buciluă et al., 2006; Hinton et al., 2015), (attention-weighted) feature maps (Romero et al., 2015; Zagoruyko & Komodakis, 2016), or latent representations (Ahn et al., 2019; Park et al., 2019; Tian et al., 2019; Laskar & Kannala, 2020). Importantly, Tian et al. (2019) show that under fair comparison, basic matching via Kullback-Leibler (KL) Divergences as used in Hinton et al. (2015) performs very well, which we also find to be the case for S2SD. This is further supported in recent few-shot learning literature (Tian et al., 2020), wherein KL-distillation alongside self-distillation (by iteratively reusing the same network as a teacher for beneficial generalization and regulatory properties (Furlanello et al., 2018; Lan et al., 2018; Zhang et al., 2019b; Abnar et al., 2020; Yun et al., 2020)) in additional meta-training stages improves feature
representation strength important for generalization (Raghu et al., 2020).

Our work is closest to Zhang et al. (2019a) and Liu et al. (2020), which propose to break down a network into a cascading set of subnetworks, wherein each subsequent subnetwork builds on its predecessors. In doing so, each subnetwork is trained independently on a classification task at hand. Knowledge distillation is then applied either from the full network (Zhang et al., 2019a) acting as a teacher or via soft targets generated from a meta-learned label generator (Liu et al., 2020), to each smaller student subnetwork during the same training run to improve overall performance. In a related manner, S2SD utilizes similar concurrent, but relational self-distillation to instead encode high-dimensional sample relation context from multiple, higher-dimensional teacher embedding spaces; this is crucial to improve the generalization capabilities of a single student embedding space for zero-shot, out-of-distribution image retrieval tasks. As such, it operates orthogonally to proposals made by Zhang et al. (2019a) and Liu et al. (2020). The concurrency of the self-distillation in turn is a consequence of the novel insight that solely the dimensionality of embedding spaces can serve as meaningful teachers, as these can be spanned cheaply over a large, shared feature backbone. The novel dimensionality-based concurrent distillation also sets S2SD apart from existing knowledge distillation applications to DML, which are done in a generic manner with separate, larger teacher networks or additional training stages (Chen et al., 2018; Yu et al., 2019; Han et al., 2019; Laskar & Kannala, 2020).

3. Method

We now introduce key elements for Simultaneous Similarity-based Self-Distillation (S2SD) to improve generalization of embedding spaces by utilizing higher dimensional context. We begin with the preliminary notation and fundamentals to Deep Metric Learning (§3.1), before defining the three key elements to S2SD: Firstly, the Dual Self-Distillation (DSD) objective, which uses KL-Distillation on a concurrently learned embedding space of higher dimensionality (§3.2) to introduce important high-dimensional context into a low-dimensional embedding space during training. We then extend this to Multiscale Self-Distillation (MSD) with distillation from several different high-dimensional embedding spaces to encourage reusability in the distilled context (§3.3). Finally, we shift to self-distillation from normalized feature representations (MSDF) to counter dimensionality bottlenecks commonly encountered in DML (§3.4).

3.1. Preliminaries

DML builds on generic Metric Learning which aims to find a (parametrized) distance metric $d_\theta : \Phi \times \Phi \to \mathbb{R}$ on the feature space $\Phi \subset \mathbb{R}^d$ over images $\mathcal{X}$ that best satisfy ranking constraints usually defined over class labels $\mathcal{Y}$. This holds also for DML. However, while Metric Learning relies on a fixed feature extraction method to obtain $\Phi$, DML introduces deep neural networks to concurrently learn a feature representation. Most such DML approaches aim to learn Mahalanobis distance metrics, which cover the parametrized family of inner product metrics (Suárez et al., 2018; Chen et al., 2019). These metrics, with some restrictions (Suárez et al., 2018; Chen et al., 2019).
et al., 2018), can be reformulated as

\[
d(\phi_1, \phi_2) = \sqrt{(L(\phi_1 - \phi_2)^T L(\phi_1 - \phi_2))}
= \|L\phi_1 - L\phi_2\|_2 = \|\psi_1 - \psi_2\|_2
\]

(1)

with \(L \in \mathbb{R}^{d \times d'}\) from \(d'\)-dim. features \(\phi_i \in \Phi\) to \(d\)-dim. embeddings \(\psi_i := (f \circ \phi)(x_i) \in \Psi_f\) with embedding function \(f : \phi_i \rightarrow L\phi_i\). Importantly, this redefines the motivation behind DML as learning \(d\)-dimensional image embeddings \(\psi\) s.t. their euclidean distance \(d(\cdot, \cdot) = \|\cdot - \cdot\|_2\) is connected to semantic similarities in \(X\). This embedding-based formulation offers the significant advantage of being compatible with fast approximate similarity search methods (e.g. (Johnson et al., 2017)), allowing for large-scale applications at test time. In this work, we assume \(\Psi_f\) to be normalized to the unit hypersphere \(S_{\Psi_f}\), which is commonly done (Wu et al., 2017; Sanakoyeu et al., 2019; Liu et al., 2017; Wang & Isola, 2020) for beneficial regularizing purposes (Wu et al., 2017; Wang & Isola, 2020). For the remainder we hence set \(\Psi\) to refer to \(S_{\Psi}\).

Common approaches to learn such a representation space involve training surrogates on ranking constraints defined by class labels. Such approaches start from pair or triplet-based ranking objectives (Hadsell et al., 2006; Schroff et al., 2015), where the latter is defined as

\[
\mathcal{L}_\text{triplet} = \frac{1}{|\mathcal{B}|} \sum_{(x_i, x_j, x_k) \in \mathcal{B}} [d(\psi_i, \psi_j) - d(\psi_i, \psi_k) + m]_+
\]

(2)

with margin \(m\) and the set of available triplets \((x_i, x_j, x_k) \in \mathcal{B}\) in a mini-batch \(\mathcal{B} \subset X\), with \(y_i = y_j \neq y_k\). This can be extended with more complex ranking constraints or tuple sampling methods. We refer to Supp. B and Roth et al. (2020b) for further insights and detailed studies.

3.2. Embedding Space Self-Distillation

For the aforementioned standard DML setting, generalization performance of a learned embedding space can be linked to the utilized embedding dimensionality. However, high dimensionality results in notably higher retrieval cost on downstream applications, which limits realistically usable dimensions. In \(S2SD\), we show that high-dimensional context can be used as a teacher during the training run of the low-dimensional base or reference embedding space. As the base embedding model is also the one that is evaluated, test time retrieval costs are left unchanged.

To achieve this, we simultaneously train an additional high-dimensional auxiliary/target embedding space \(\Psi_g := (g \circ \phi)(X)\) spanned by a secondary embedding branch \(g\). \(g\) is parametrized by a MLP or a linear projection, similar to the base embedding space \(\Psi_f\) spanned by \(f\), see §3.1. Both \(f\) and \(g\) operate on the same large, shared feature backbone \(\phi\). For simplicity, we train \(\Psi_f\) and \(\Psi_g\) using the same DML objective \(\mathcal{L}_{\text{DML}}\).

Unfortunately, higher expressivity and improved generalization of high-dimensional embeddings in \(\Psi_g\) hardly benefit the base embedding space, even with a shared feature backbone. To explicitly leverage high-dimensional context for our base embedding space, we utilize knowledge distillation from target to base space. However, while common knowledge distillation approaches match single embeddings or features between student and teacher, the different dimensionalities in \(\Psi_f\) and \(\Psi_g\) inhibit naive matching.

Instead, \(S2SD\) matches sample relations (see e.g. (Tian et al., 2019)) defined over batch-similarity matrices \(D \in \mathbb{R}^{B \times B}\) in base and target space, \(D^f\) and \(D^g\), with batchsize \(B\). We thus encourage the base embedding space to relate different samples in a similar manner to the target space. To compute \(D\), we use a cosine similarity by default, given as \(D_{i,j} = \psi_i^T \psi_j\), since \(\psi_i\) is normalized to the unit hypersphere. Defining \(\sigma_{\text{max}}\) as the softmax operation and \(D_{\text{KL}}(p, q) = \sum \log(p) \log(p) / \log(q)\) as the Kullback-Leibler divergence, we thus define the simultaneous self-distillation objective as

\[
\mathcal{L}_{\text{dist}}(D^f, D^g) = \sum_{i}^{B} D_{\text{KL}} \left( \sigma_{\text{max}} \left( D^f_i / T \right) , \sigma_{\text{max}} \left( D^g_i / T \right) \right)
\]

(3)

with temperature \(T\), as visualized in Figure 1. \((^\dagger)\) denotes no gradient flow to target branches \(g\) as we only want the base space to learn from the target space. By default, we match rows or columns of \(D, D_{\dagger}\) effectively distilling the relation of an anchor embedding \(\psi_i\) to all other batch samples. Embedding all batch samples in base dimension, \(\Psi^B_f : B \leftrightarrow \psi_f(B)\), and higher dimension, \(\Psi^B_g : B \leftrightarrow \psi_g(B)\), the (simultaneous) Dual Self-Distillation (DSD) training objective then becomes

\[
\mathcal{L}_{\text{DSD}}(\Psi^B_f, \Psi^B_g) = \frac{1}{2} \cdot \left[ \mathcal{L}_{\text{DML}}(\Psi^B_f) + \mathcal{L}_{\text{DML}}(\Psi^B_g) \right] + \gamma \cdot \mathcal{L}_{\text{dist}}(D^f, D^g)
\]

(4)

3.3. Reusable Relations by Multiscale Distillation

While DSD encourages the reference embedding space to recover complex sample relations by distilling from a higher-dimensional target space spanned by \(g\), it is not known \(a\ priori\) which distillable sample relations actually benefit generalization of the reference space.

To encourage the usage of sample relations that more likely aid generalization, we follow insights made in Raghu et al. (2020) on the connection between \(\text{reusability}\) of features across multiple tasks and better generalization thereof. We motivate reusability in \(S2SD\) by extending DSD to Multiscale Self-Distillation (MSD) with distillation instead from \(m\) multiple different target spaces spanned
by \( G = \{g_k\}_{k \in \{1, m\}} \). Importantly, each of these high-dimensional target spaces operate on different dimensionalities, i.e. \( \text{dim } f < \text{dim } g_1 < \ldots < \text{dim } g_{m-1} < \text{dim } g_m \).

As this results in each target embedding space encoding sample relations differently, application of distillation across all spaces spanned by \( G \) pushes the base branch towards learning from sample relations that are reusable across all higher dimensional embedding spaces and thereby more likely to generalize (see also Fig. 1).

Specifically, given the set of target similarity matrices \( \{D^k\}_{k \in \{f, g_1, \ldots, g_m\}} \) and target batch embeddings \( \{\Psi^B_k\}_{k \in \{f, g_1, \ldots, g_m\}} \), we then define the MSD training objective as

\[
\mathcal{L}_{\text{MSD}}(\Gamma^m) = \frac{1}{2} \left[ \mathcal{L}_{\text{DML}}(\Psi^B_f) + \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\text{DML}}(\Psi^B_{g_i}) \right] + \frac{\gamma}{m} \sum_{i=1}^{m} \mathcal{L}_{\text{dist}}(D^f, D^{g_i})
\]

(5)

3.4. Tackling the Dimensionality Bottleneck by Feature Space Self-Distillation

As noted in §3.1, the base embedding \( \Psi \) utilizes linear projections \( f \) from the (penultimate) feature space \( \Phi \) where \( \text{dim } \Phi \) is commonly much larger than \( \text{dim } \Psi \). While compressed semantic spaces encourage stronger representations (Alemi et al., 2016; Dai & Wipf, 2019) to be learned, Milbich et al. (2020) show that the actual test performance of the lower-dimensional embedding space \( \Psi \) is inferior to that of the non-adapted, but higher-dimensional feature space \( \Phi \).

This supports a dimensionality-based loss of information beneficial to generalization, which can hinder the base embedding space to optimally approximate the high-dimensional context introduced in §3.2 and 3.3.

To rectify this, we apply self-distillation following eq. 3 on the normalized feature representations \( \Phi^\phi \) generated by normalizing the backbone output \( \phi \). With the batch of normalized feature representations \( \Psi^B_{\Phi^\phi} \) we get multiscale self-distillation with feature distillation (MSDF) (see also Fig. 1)

\[
\mathcal{L}_{\text{MSDF}}(\Gamma^m, \Psi^B_{\Phi^\phi}) = \mathcal{L}_{\text{MSD}}(\Gamma^m) + \gamma \mathcal{L}_{\text{dist}}(D^f, D^{\Phi^\phi})
\]

(6)

In the same manner, one can also address other architectural information bottlenecks such as through the generation of feature representations from a single global pooling operation. While not noted in the original publication, Kim et al. (2020) address this in the official code release by using both global max- and average pooling to create their base embedding space. While this naive usage changes the architecture at test time, in S2SD we can fairly leverage potential benefits by only spanning the auxiliary spaces (and distilling) from such feature representations (denoted as DSDA/MSDA/MSDFA).

4. Experimental Setup

We study S2SD in four experiments to establish 1) method ablation performance & relative improvements, 2) state-of-the-art, 3) comparisons to standard 2-stage distillation, benefits to low-dimensional embedding spaces & generalization properties and 4) motivation for architectural choices.

Method Notation. We abbreviate ablations of S2SD (see §3) in our experiments as: DSD & MSD for Dual (3.2) & Multiscale Self-Distillation (3.3). MSDF the addition of Feature distillation (3.4) and DSDA/MSDA(F)A the inclusion of multiple pooling operations in the auxiliary branches (also §3.4).

4.1. Experiments

Fair Evaluation of S2SD. §5.1 specifically applies S2SD and its ablations to three DML baselines. To show realistic benefit, S2SD is applied to best-performing objectives evaluated in Roth et al. (2020b), namely (i) Margin loss with Distance-based Sampling (Wu et al., 2017), (ii) their proposed Regularized Margin loss and (iii) Multisimilarity loss (Wang et al., 2019), following their experimental training pipeline. This setup utilizes no learning rate scheduling and fixes common implementational factors of variation in DML pipelines such as batchsize, base embedding dimension, weight decay or feature backbone architectures to ensure comparability in DML (more details in Supp. A.2).

As such, our results are directly comparable to their large set of examined methods and guaranteed that relative improvements solely stem from the application of S2SD.

Comparison to literature. §5.2 further highlights the benefits of S2SD by comparing S2SD’s boosting properties across literature standards, with different backbone architectures and base embedding dimensions: (i) ResNet50 with \( d = 128 \) (Wu et al., 2017; Roth et al., 2019) and (ii) \( d = 512 \) (Zhai & Wu, 2018) as well as (3) variants to Inception-V1 with Batch-Normalization at \( d = 512 \) (Wang et al., 2019; Qian et al., 2019; Milbich et al., 2020). Only here do we conservatively apply learning rate scheduling, since all references noted in Table 2 employ scheduling as well. We categorize published work based on backbone architecture and embedding dimension for fairer comparison. Note that this is a less robust comparison than done in §5.1, due to potential implementation differences between our pipeline and reported literature results.

Comparison to 2-Stage Distillation. §5.3 compares S2SD to 2-stage distillation, investigates benefits to very low dimensional reference spaces and examines the connection between improvements and changes in embedding space density and spectral decay (see Supp. D), which have been linked to improved generalization.

Ablation Study. §5.4 ablates and motivates architectural
Table 1. S2SD comparison against strong baseline objectives. Bold denotes best results per objective, bluebold marks best overall results. mAP@R results as proposed in (Roth et al., 2020b) and (Musgrave et al., 2020) as well as ProxyAnchor evaluations (Kim et al. (2020), using a different setup) can be found in the Supplementary (Table 5 and 6), further showing the notable benefits of S2SD.

<table>
<thead>
<tr>
<th>Benchmarks →</th>
<th>CUB200-2011</th>
<th>CARs196</th>
<th>SOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approaches ↓</td>
<td>R@1</td>
<td>NMI</td>
<td>R@1</td>
</tr>
<tr>
<td>Margin, $\beta = 1.2$, (Wu et al., 2017)</td>
<td>63.19 ± 0.46</td>
<td>68.21 ± 0.33</td>
<td>79.86 ± 0.33</td>
</tr>
<tr>
<td>+ DSD</td>
<td>65.11 ± 0.18</td>
<td>69.65 ± 0.44</td>
<td>83.19 ± 0.18</td>
</tr>
<tr>
<td>+ MSD</td>
<td>66.13 ± 0.34</td>
<td>70.83 ± 0.27</td>
<td>83.63 ± 0.31</td>
</tr>
<tr>
<td>+ MSDF</td>
<td><strong>67.58 ± 0.32</strong></td>
<td><strong>71.47 ± 0.19</strong></td>
<td><strong>85.55 ± 0.23</strong></td>
</tr>
<tr>
<td>+ MSDFA</td>
<td>67.21 ± 0.23</td>
<td>71.43 ± 0.25</td>
<td><strong>86.45 ± 0.35</strong></td>
</tr>
<tr>
<td>R-Margin, $\beta = 0.6$, (Roth et al., 2020b)</td>
<td>64.93 ± 0.42</td>
<td>68.36 ± 0.32</td>
<td>82.37 ± 0.13</td>
</tr>
<tr>
<td>+ DSD</td>
<td>66.58 ± 0.08</td>
<td>70.03 ± 0.41</td>
<td>84.64 ± 0.16</td>
</tr>
<tr>
<td>+ MSD</td>
<td>68.12 ± 0.27</td>
<td>70.67 ± 0.16</td>
<td>85.01 ± 0.10</td>
</tr>
<tr>
<td>+ MSDF</td>
<td>68.12 ± 0.30</td>
<td><strong>71.80 ± 0.33</strong></td>
<td>85.78 ± 0.22</td>
</tr>
<tr>
<td>+ MSDFA</td>
<td><strong>68.58 ± 0.26</strong></td>
<td>71.64 ± 0.40</td>
<td><strong>86.81 ± 0.35</strong></td>
</tr>
<tr>
<td>Multisimilarity (Wang et al., 2019)</td>
<td>62.80 ± 0.70</td>
<td>68.55 ± 0.38</td>
<td>81.68 ± 0.19</td>
</tr>
<tr>
<td>+ DSD</td>
<td>65.57 ± 0.26</td>
<td>70.08 ± 0.33</td>
<td>83.51 ± 0.20</td>
</tr>
<tr>
<td>+ MSD</td>
<td>65.80 ± 0.16</td>
<td>70.66 ± 0.01</td>
<td>83.95 ± 0.10</td>
</tr>
<tr>
<td>+ MSDF</td>
<td>67.04 ± 0.29</td>
<td><strong>71.87 ± 0.19</strong></td>
<td>85.69 ± 0.19</td>
</tr>
<tr>
<td>+ MSDFA</td>
<td><strong>67.68 ± 0.29</strong></td>
<td>71.40 ± 0.21</td>
<td><strong>85.89 ± 0.15</strong></td>
</tr>
</tbody>
</table>

choices in S2SD used throughout §4. Pseudo code and detailed results are available in Supp. F, G, and I.

4.2. Implementation

Datasets & Evaluation. In all experiments, we evaluate on standard DML benchmarks: CUB200-2011 (Wah et al., 2011), CARs196 (Krause et al., 2013) and Stanford Online Products (SOP) (Oh Song et al., 2016). Performance is measured in recall at 1 (R@1) and at 2 (R@2) (Jegou et al., 2011) as well as Normalized Mutual Information (NMI) (Manning et al., 2010). Results measured on mean Average Precision evaluated on Recall (mAP@R) are available in the Supplementary along with additional dataset details.

Experimental Details. Our implementation follows Roth et al. (2020b) for comparability, using frozen Batch-Normalization (Ioffe & Szegedy, 2015) whenever a ResNet50 (He et al., 2016) is utilized, Adam (Kingma & Ba, 2015) with learning rate $10^{-5}$ for training and weight decay (Krogh & Hertz, 1992) of $4 \cdot 10^{-5}$ for regularization. Additional details are available in Supp. (A). For §5.1-5.4, we only adjust the respective pipeline elements in questions. For S2SD, unless noted otherwise (s.a. in §5.4), we set $\gamma = 50, T = 1$ for all objectives on CUB200 and CARs196, and $\gamma = 5, T = 1$ on SOP. DSD uses target-dim. $d = 2048$ and MDF uses target-dims. $d \in [512, 1024, 1536, 2048]$. We found it beneficial to activate the feature distillation after $n = 1000$ iterations to ensure that meaningful features are learned first before feature distillation is applied. Additional embedding spaces are spanned by two layer MLPs with row-wise KL-distillation of high-dimensional similarities (eq. 3), applied as in $\mathcal{L}_{\text{mul}}$ (eq. 5). By default, we use Multisimilarity Loss as stand-in for $\mathcal{L}_{\text{DML}}$. Hyperparameters were determined previous to the result runs using a 80-20 training and validation split, similar to Roth et al. (2020b) and Kim et al. (2020).

5. Results

5.1. Fair performance study

In Tab. 1 (full table in Supp. Tab. 4), we show that under fair experimental protocol, utilizing S2SD and its objectives gives an objective and benchmark independent, significant boost in performance by up to 7% opposing the existing DML objective performance plateau. This holds even for previous state-of-the-art regularized objectives s.a. R-Margin loss as well as proxy-based objectives such as ProxyAnchor (Kim et al. (2020), see Supplementary), highlighting the effectiveness of S2SD for DML. Across objectives, S2SD-based changes in wall-time do not exceed negligible 5% with only minor convergence impacts.

5.2. Setting a new State-of-the-Art

Motivated by Tab. 1, we use MSDFA for CUB200/CARS196 and MSDF for SOP. Table 2 shows that S2SD can boost baseline objectives to reach and even surpass other state-of-the-
art methods, in parts with a notable margin. This holds even when compared to much more complex methods with feature mining or RL-policies such as MIC (Roth et al., 2019), DiVA (Milbich et al., 2020) or PADS (Roth et al., 2020a), to which S2SD operates orthogonally. Finally, we note that these insights are true even with our results reported with confidence intervals, which is commonly neglected in DML.

5.3. Benefits of S2SD

Comparison to standard distillation. With a student $S$ using the same objective and embedding dimensionality as the reference branch in DSD) and a teacher $T$ at the highest optimal dimensionality $d = 2048$, we find that separating DSD into a standard 2-stage distillation setup actually degenerates performance (see Fig. 3A, compare to Dist.). In addition, S2SD allows for easy integration of teacher ensembles, realized by MSD($FA$), to even outperform the teacher by a notable margin. This is specifically interesting as S2SD retains the operating embedding dimensionality of the student.

Benefits to lower base dimensions. We now show that our module is able to vastly boost networks limited to very low embedding dimensions, which we visualize in figure 3B). For example, networks operating on $d = 32$ & 64 trained with S2SD can match the performance of networks trained and evaluated on embedding dimensions four or eight times the size. For $d = 128$, S2SD even outperforms the highest dimensional baseline at $d = 2048$ by a large margin.

Embedding space metrics. We now look at relative changes in embedding space density and spectral decay (see supplementary of Roth et al. (2020b)) when applying S2SD. Our study, visualized in figure 2, shows that the application of S2SD increases embedding space density and lowers the spectral decay (thus providing a more feature-diverse embedding space) across criteria, which is aligned with properties of improved generalization in DML as noted in Roth et al. (2020b).

5.4. Motivating S2SD Architecture Choices

Distillation improves generalization through S2SD. Fig. 3A (Joint) and Fig. 3F ($\gamma = 0$) highlight how crucial self-distillation is, as incorporating a secondary embedding space without any distillation link hardly improves performance. Fig. 3A (Concur.) further shows that joint training of a detached reference embedding $f$, while otherwise training in high dimension, similarly doesn’t offer notable improvement. Finally, Figure 3F shows robustness to changes in $\gamma$, with peaks around $\gamma = 50$ and $\gamma = 5$ for CUB200/CARS196 and SOP. We also found best performance $T \in [0.2, 2]$ and hence set $T = 1$
Simultaneous Similarity-based Self-Distillation for Deep Metric Learning

Figure 3. S2SD study and ablations. (A) DSD outperforms comparable two-stage distillation on student S (Dist.) using teacher (T), with MSD(FA) even outperforming T. We further see that distillation is essential - training multiple spaces in parallel (Joint.) or a detached lower-dimensional base embedding (Concur.) gives little benefit. (B) We see benefits across base dimensionalities, especially in the low-dimensional regime. (C) We find KL-distillation between similarity vectors (R-KL) to work best. (D) An additional non-linearity in aux. branches gives a boost, but going deeper hurts generalization. (E) Distilling each aux. embed. space (Multi) separately compares favourable against other distillation setups s.a. Nested and Chained distillation. (F) Performance is robust to changes in weight values.

by default.

Best way to enforce reusability. To motivate our many-to-one self-distillation $\mathcal{L}_{\text{MSD}}$ (eq. 5, here also dubbed $\mathcal{L}_{\text{Multi}}$), we evaluate against other distillation setups that could support reusability of distilled sample relations: (1) Nested distillation, where instead of distilling all target spaces only to the reference space, we distill from a target space to all lower-dimensional embedding spaces:

$\mathcal{L}_{\text{Nested}}(\Gamma^m) = \frac{1}{2} \mathcal{L}_{\text{DML}}(\Psi^B_f) + \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\text{DML}}(\Psi^B_{g_i})$ 

$$\quad + \frac{\gamma}{m-1} \sum_{i=0, j=1, j \neq i}^{m} \mathcal{L}_{\text{dist}}(\Psi^B_{g_i}, \Psi^B_{g_j})$$

(7)

In the second term, $g_0$ denotes the base embedding $f$.

(2) Chained distillation, which distills target spaces only to the immediate lower-dimensional embedding space:

$\mathcal{L}_{\text{Chained}}(\Gamma^m) = \frac{1}{2} \mathcal{L}_{\text{DML}}(\Psi^B_f) + \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}_{\text{DML}}(\Psi^B_{g_i})$

$$\quad + \frac{\gamma}{m} \sum_{i=1}^{m-1} \mathcal{L}_{\text{dist}}(\Psi^B_{g_i}, \Psi^B_{g_{i+1}})$$

(8)

Figure 3E shows that a many-to-one distillation performs notably better, supporting the reusability aspect and $\mathcal{L}_{\text{Multi}}$ as our default method.

Choice of distillation method & branch structures. Fig. 3C evaluates various distillation objectives, finding KL-divergence between vectors of similarities to perform better than KL-divergence applied over full similarity matrices or row-wise means thereof, as well as cosine/euclidean distance-based distillation (see e.g. Yu et al. (2019)). Figure 3D shows insights into optimal auxiliary branch structures, with two-layer MLPs giving the largest benefit, although even a linear target mapping reliably boosts performance. This coincides with insights made by Chen et al. (2020). Further network depth only deteriorates performance.

6. Conclusion

In this paper, we propose a novel DML training paradigm based on dimensionality-based knowledge distillation, Simultaneous Similarity-based Self-Distillation (S2SD). S2SD allows for the inclusion of reusable, context-rich, high-dimensional relational information for improved generalization. This is achieved by solving the standard DML objective simultaneously in higher-dimensional embedding spaces while applying knowledge distillation concurrently between these high-dimensional teacher spaces and a lower-dimensional reference space. In doing so, S2SD introduces little additional computational overhead, with no extra cost at test time. Thorough ablations and experiments show S2SD significantly improving the generalization performance of existing DML objectives regardless of embedding dimensionality, thereby also setting a new state-of-the-art.
Acknowledgements

We would like to thank Samarth Sinha (University of Toronto, Vector), Matthew McDermott (MIT) and Mengye Ren (University of Toronto, Vector) for insightful discussions and feedback on the paper draft. This work was funded in part by a CIFAR AI Chair at the Vector Institute, Microsoft Research, and an NSERC Discovery Grant. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute www.vectorinstitute.ai/#partners.

Reviewer Comments

Re: More advanced baseline methods. We do believe that our reported results are representative of the current state of Deep Metric Learning - under fair comparison (Roth et al., 2020b), Multisimilarity and Margin loss achieve best or competitive results. In addition, S2SD is applied to regularized Deep Metric Learning (R-Margin loss (Roth et al., 2020b)) and shows high improvements throughout. ProxyAnchor (Kim et al., 2020) was not included in the literature. Regarding general convergence behaviour and training parameters. We offer a dimensionality versus synthetic data mining international conference on Knowledge discovery and data mining (KDD), pp. 535–541, 2006.


Simultaneous Similarity-based Self-Distillation for Deep Metric Learning


Simultaneous Similarity-based Self-Distillation for Deep Metric Learning


Simultaneous Similarity-based Self-Distillation for Deep Metric Learning


