Self-Taught Metric Learning without Labels —Supplementary Material—

Sungyeon Kim[1] Dongwon Kim[1] Minsu Cho[1,2] Suha Kwak[1,2]
Dept. of CSE, POSTECH[1] Graduate School of AI, POSTECH[2]
{sungyeon.kim, kdwon, mscho, suha.kwak}@postech.ac.kr
http://cvlab.postech.ac.kr/research/STML/

This supplementary material provides further analyses, societal impacts, and additional experimental results, all of which are left out from the main paper due to the space limit. In Section 1, we first discuss properties of the contextualized semantic similarity in detail. Finally, in Section 2, we present t-SNE visualizations of the learned embedding space and qualitative examples of image retrieval on the three benchmark datasets.

1. In-depth Analysis on Contextualized Semantic Similarity

STML employs the contextualized semantic similarity as pseudo label that can capture both inter- and intra-class relations. To better understand its property, we illustrate in Figure 1 how the contextualized semantic similarity is determined in four different cases of relations between samples. As shown in Figure 1(b), when a pair of samples not only share many k-reciprocal nearest neighbors but also have a small Euclidean distance, they obviously have high contextualized semantic similarity; they are visually similar and placed on the same manifold, thus highly likely to belong to the same class. In contrast, as shown in Figure 1(c), two samples that are far apart and share no neighbor have a low contextualized semantic similarity as they are semantically unrelated; we found that relations of most pairs of unlabeled data fall into this case. In Figure 1(a) and 1(d), where only one of the pairwise and contextual similarities is high, two samples have a higher contextualized semantic similarity than that in the case of Figure 1(c), but lower than that in the case of Figure 1(b). Note that although the class-equivalence between samples is uncertain in the cases of Figure 1(a) and 1(d), the pairwise and contextual similarities could be overly high in each of these cases, leading to noisy synthetic supervision. On the other hand, the contextualized semantic similarity is modest thus provides reliable pseudo labels in these cases.

The contextualized semantic similarity allows the relaxed contrastive loss to exploit reliable supervision. To demonstrate this, we empirically analyze the information provided by the contextualized semantic similarity. Figure 2 presents the top-8 samples in a mini-batch of each query in terms of their contextualized semantic similarity. The results are obtained by the source embedding model where the number of nearest neighbor k is 5. As shown in the figure, the contextualized semantic similarity is highly correlated with the groundtruth class-equivalence on all datasets. In particular, birds of the same class are assigned higher contextualized semantic similarities although all birds are floating on water with similar poses in the 2nd row of Figure 2. Moreover, as can be seen in the 1st and 6th rows of Figure 2, the contextualized semantic similarity successfully captures class-equivalence even under view-point variations.

2. Additional Qualitative Results

We present t-SNE visualizations of the embedding space learned by the framework at every 30 epochs in Figure 5. At the beginning of learning, embedding vectors of the same class are spread out while overlapping with those of other classes. As the epoch goes on, relevant embedding vectors are gradually grouped, and embedding vectors of same class are aggregated in same cluster at the end of learning. In addition, more qualitative retrieval results of our model at every 30 epochs of training on the CUB, Cars and SOP datasets are presented in Figure 6, Figure 7, and Figure 8, respectively. As training progresses, the model trained by STML distinguishes samples of the same class more accurately among visually similar samples.

All the results in these figures are obtained from unseen class samples without any manual annotation. The resutls suggest that STML allows the final embedding model to generalize well even to unseen classes through reliable pseudo labels considering intra- and inter-class relations.

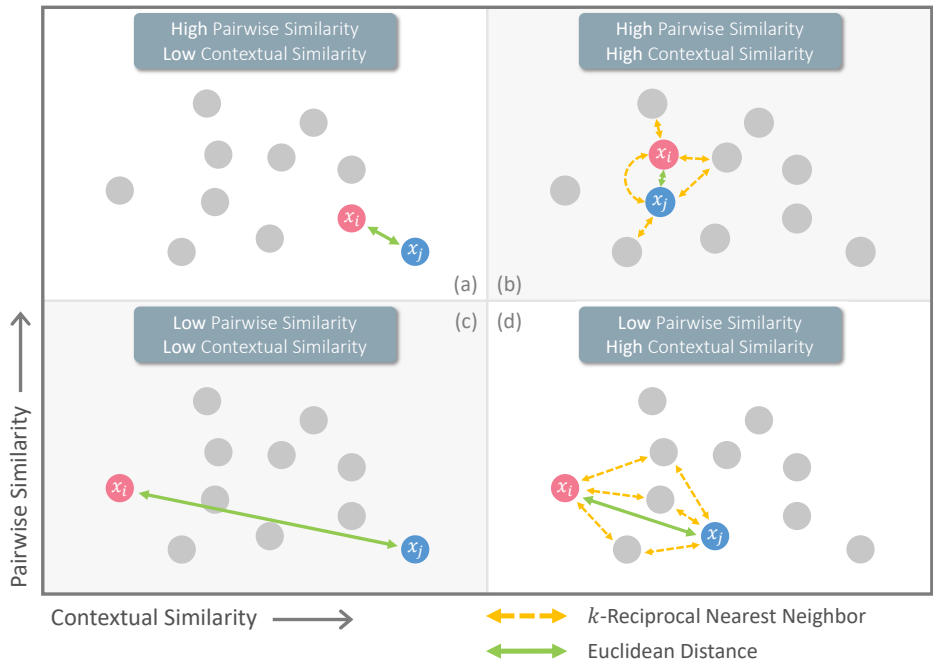

Figure 1. Difference in contextualized semantic similarity according to the relation between two samples.

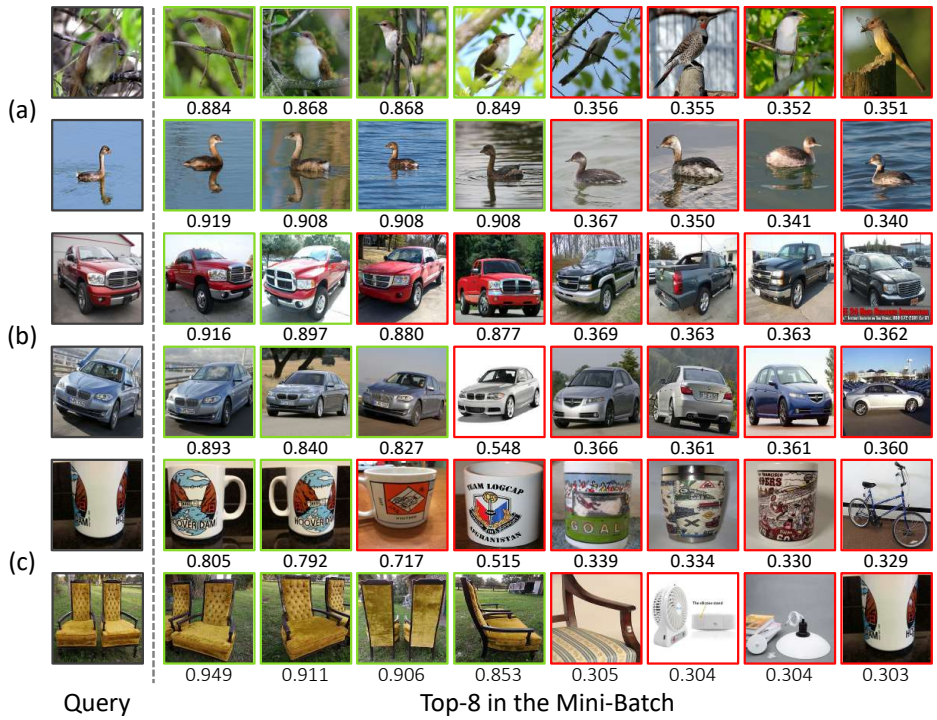

Figure 2. Image pairs sorted by their contextualized semantic similarity on (a) CUB-200-2011, (b) Cars-196, and (c) SOP datasets. Images with green boundary are of the same class as query and those with red boundary are of a different class from query.

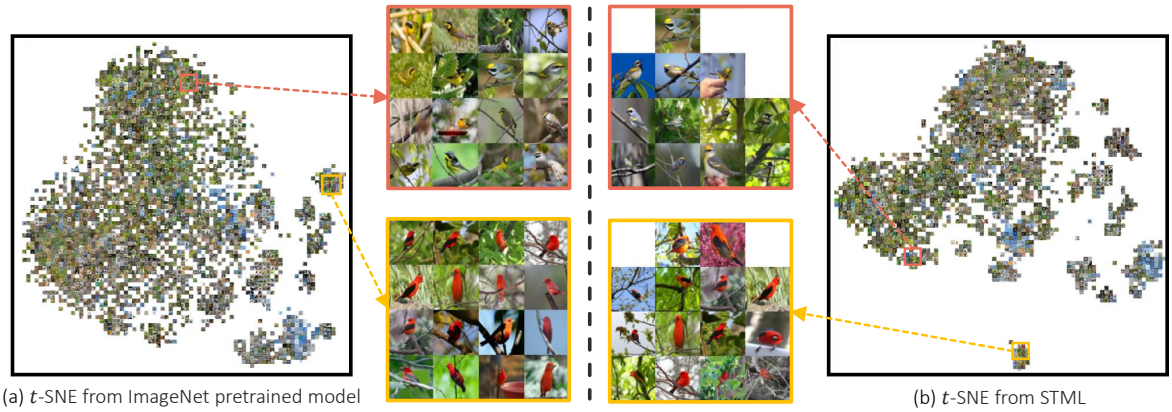

Figure 3. t-SNE visualization of our model at every 30 epochs on the (a) CUB-200-2011, (b) Cars-196, and (C) SOP datasets. Each color indicates distinct classes. For visualization, 20 classes randomly selected from the test set are used in (a) and (b), and 60 classes are used in (c).

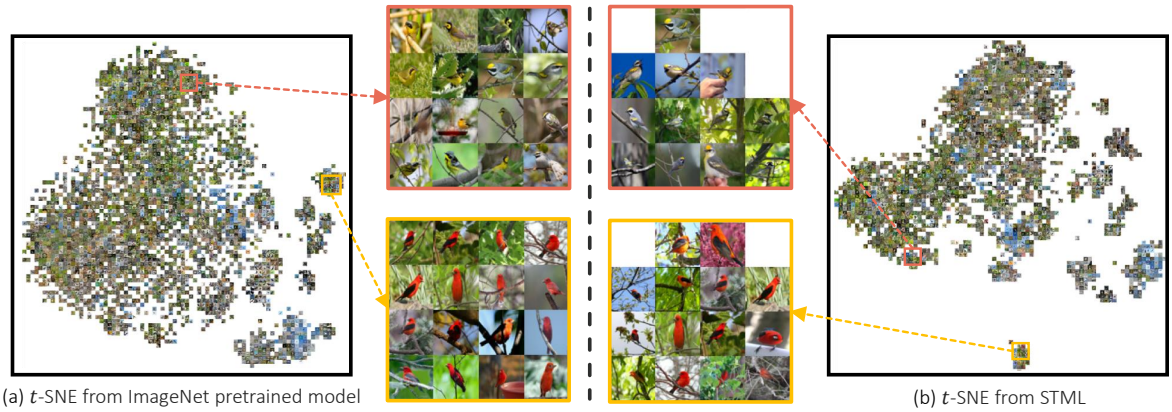

Figure 4. t-SNE visualization of our model at every 30 epochs on the (a) CUB-200-2011, (b) Cars-196, and (C) SOP datasets. Each color indicates distinct classes. For visualization, 20 classes randomly selected from the test set are used in (a) and (b), and 60 classes are used in (c).

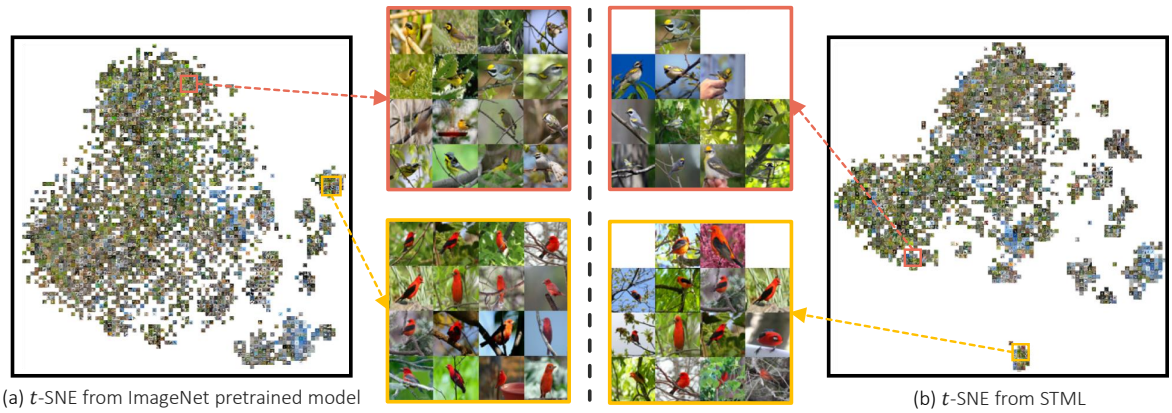

Figure 5. t-SNE visualization of our model at every 30 epochs on the (a) CUB-200-2011, (b) Cars-196, and (C) SOP datasets. Each color indicates distinct classes. For visualization, 20 classes randomly selected from the test set are used in (a) and (b), and 60 classes are used in (c).

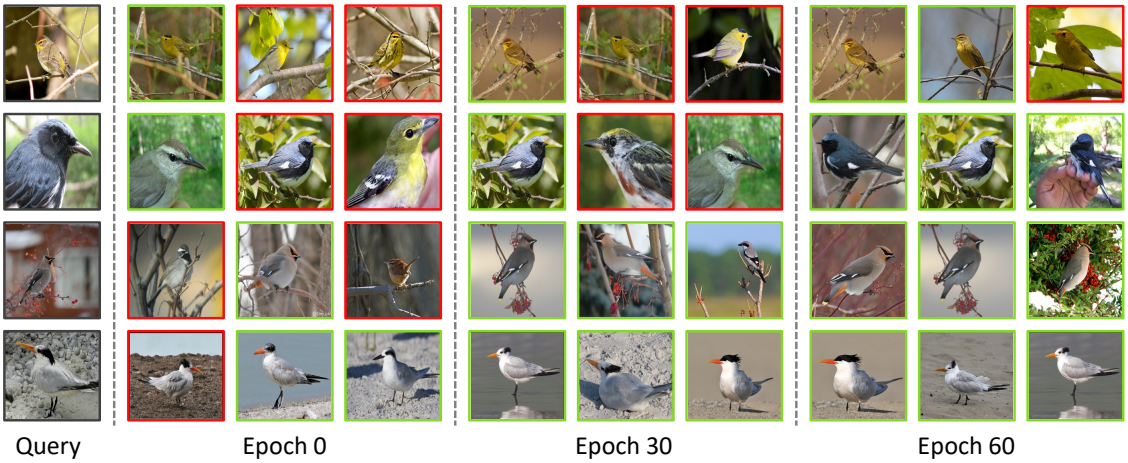

Figure 6. Top-3 retrievals of our model at every 30 epochs on the CUB-200-2011 datasets. Images with green boundary are correct and those with red boundary are incorrect.

Figure 7. Top-3 retrievals of our model at every 30 epochs on the Cars-196 datasets. Images with green boundary are correct and those with red boundary are incorrect.

Figure 8. Top-3 retrievals of our model at every 30 epochs on the SOP datasets. Images with green boundary are correct and those with red boundary are incorrect.