Contrastive Loss based on Contextual Similarity for Image Classification: Supplemental Document

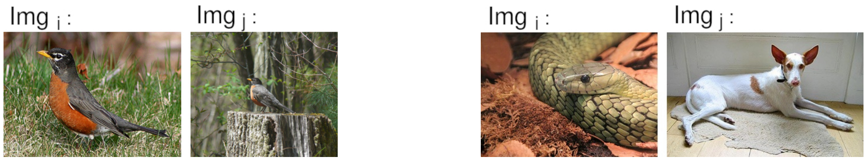
1. CODE AND IMPLEMENTATION

The code of the proposed approach is based on the Supervised Contrastive Loss (SupCon) [1] implementation: https://github.com/HobbitLong/SupContrast. The code of our proposed approach (CCL) is available here: https://github.com/lucasPV/CCL.

2. ADDITIONAL RESULTS AND VISUALIZATIONS

A. Bidimensional Spaces

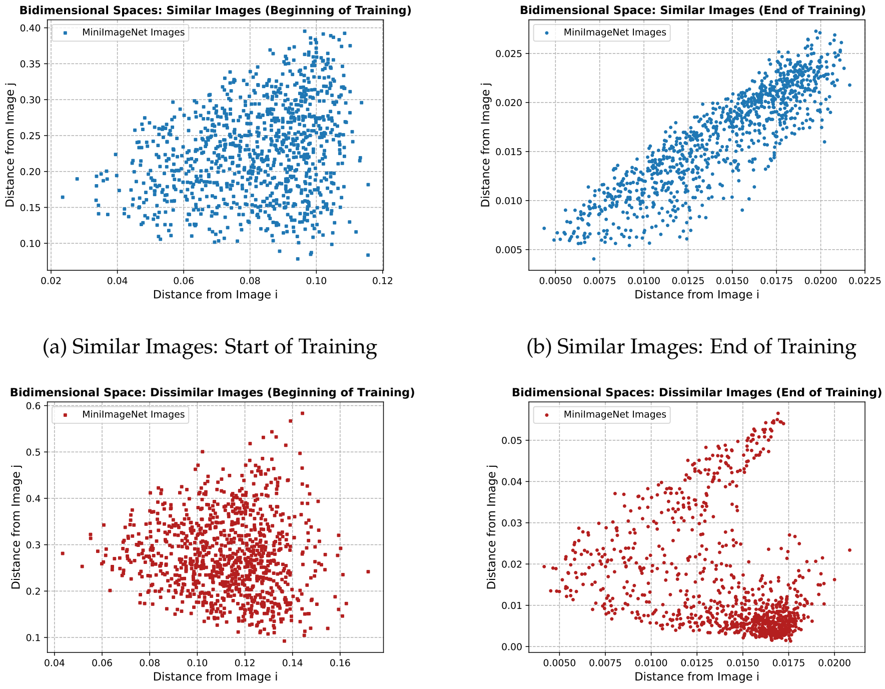
To illustrate the proposed loss, a visualization is created to show the distribution of elements by their distances at the beginning and end of the training process. Figure S1 presents the pairs of images considered as references. Figure S2 shows the plots for the same (blue) and different (red) classes. Each plot contains 1000 dots, which correspond to the top-1000 nearest neighbors of img_i. Each dot represents a distinct image, and its position is determined based on the distance from the reference images.



(a) Similar reference images

(b) Dissimilar reference images

Fig. S1. MiniImageNet [2] dataset images used as references for the bi-dimensional space plots.



(a) Similar Images: Start of Training

(b) Similar Images: End of Training

(c) Dissimilar Images: Start of Training

(d) Dissimilar Images: End of Training

Fig. S2. Bidimensional space for similar and dissimilar images on the MiniImageNet dataset at the start (10 epochs) and end (300 epochs) of training using the proposed CCL.

Initially, the distributions are completely chaotic as shown in *(a)* and *(c)*. Notice that as training enhances the separability between classes, in *(b)*, the dots tend to align in a line from bottom to top, left to right. Conversely, in *(d)*, they tend to form a line from bottom to top, right to left.

B. Dynamic Neighborhood Size (k)

During the training process, we adopted a parameter k for the neighborhood size considering a logarithmic decay across epochs described in the paper. The optimal value of k tends to vary throughout the training process. At the beginning of the training, the model uses larger values of k to contract larger chunks of data in the embedding space. In contrast, at the end of the training, the model should use smaller neighborhood sizes to ensure smaller adjustments of the model weights. To help understand the effect of k on neighborhood formation, Figure S3 illustrates a binary classification scenario where k decreases during training. Additionally, Figure S4 presents three curves considering the logarithmic k decay proposed for different numbers of epochs (100, 200, 300) and the k_{start} used for each.

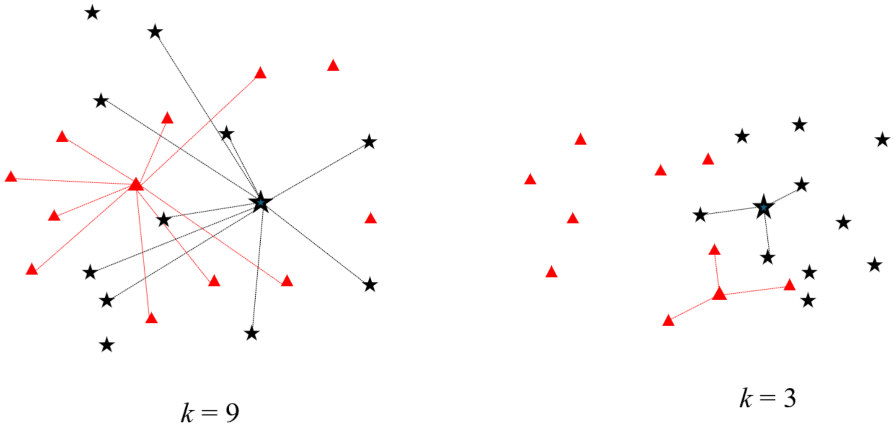


Fig. S3. Example of binary classification illustrating the effect of decreasing k from a larger value at the beginning of training ($k = 9$) to a lower value at the end of training ($k = 3$).

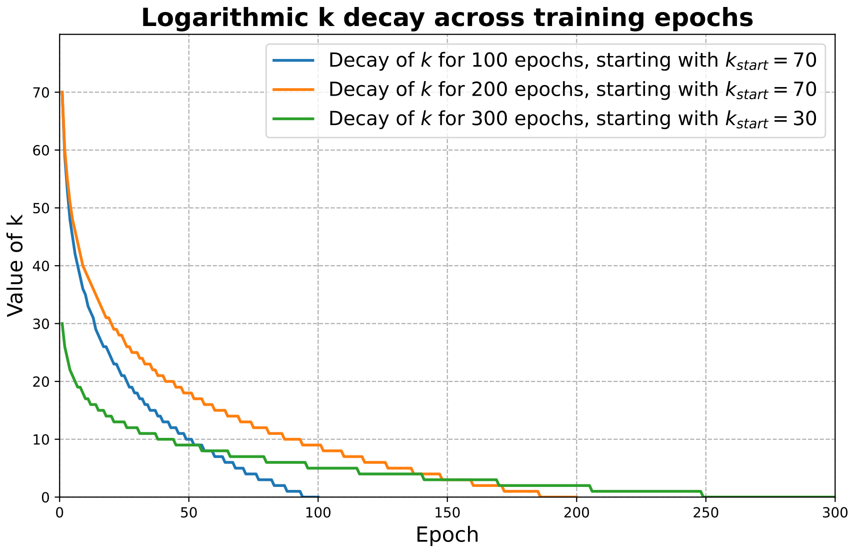


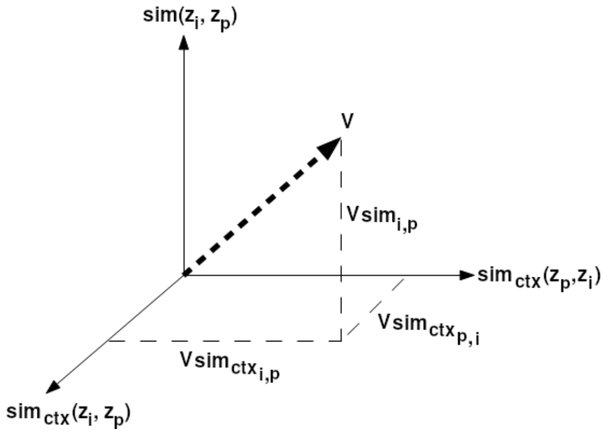
Fig. S4. Logarithmic k decay curves for different numbers of epochs (100, 200, 300) and their respective initial k_{start} values.

C. Contextual Contrastive Similarity (sim_{ccl})

The length of a vector $V = (a, b, c)$ in a 3D space is given by $\sqrt{a^2 + b^2 + c^2}$. As described in the paper, we propose the contextual contrastive similarity between i and p as follows:

$$\text{sim}_{\text{ccl}}\left(z_i, z_p, k\right) = \sqrt{\text{sim}\left(z_i, z_p\right)^2 + \text{sim}_{\text{ctx}}\left(z_i, z_p, k\right)^2 + \text{sim}_{\text{ctx}}\left(z_p, z_i, k\right)^2}.$$

Figure S5 illustrates that sim_{ccl} can be understood as length of a vector V: $|V| = \text{sim}_{\text{ccl}}\left(z_i, z_p\right)$.



$\text{sim}(z_i, z_p)$
V
$V\text{sim}_{i,p}$
$\text{sim}_{\text{ctx}}(z_p, z_i)$
$V\text{sim}_{\text{ctx}_{p,i}}$
$V\text{sim}_{\text{ctx}_{i,p}}$
$\text{sim}_{\text{ctx}}(z_i, z_p)$

Fig. S5. Illustration of a tridimensional space, where each component is used to compute the sim_{ccl} as the norm of a vector V.

D. Gains of the Proposed Approach

The experimental evaluation showed that the CCL results are consistently better than SupCon [1] and SimCLR [3], another method commonly used as a baseline in this task. Figure S6 presents a plot that evinces the capacity of CCL to provide gains when compared to SupCon for three datasets and with higher values as the training set size decreases. The integration of contextual information within the contrastive loss significantly improved the results, as initially hypothesized, with gains up to 10.759%.

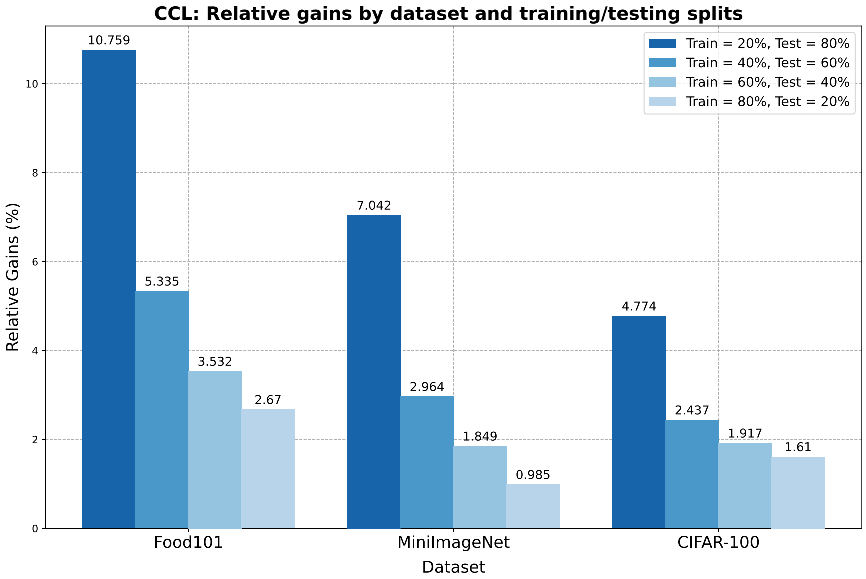
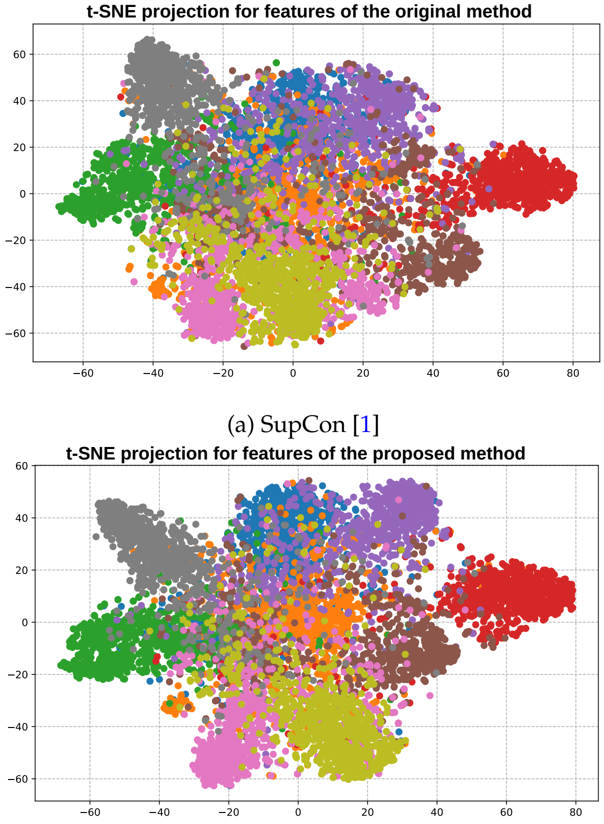


Fig. S6. Relative gains (%) obtained by CCL in comparison to SupCon for different train/test splits considering 100 training epochs.

E. t-SNE Projection

Finally, we present a qualitative result using t-SNE [4] projection on the Food101 [5] dataset considering 9 random classes. Figure S7 presents the results considering the features from the SupCon and the CCL losses, which were extracted from the linear classification model on the test set. Each color represents a different class. Notice that our approach, shown in plot *(b)*, presents better class separability. In plot *(a)*, for example, the orange group is barely visible, and groups yellow and pink overlap significantly. Additionally, other groups, such as the red and gray, are closer to the others. All these cases were improved in plot *(b)*.



(a) SupCon [1]

(b) CCL (ours)

Fig. S7. t-SNE visualization for 9 classes comparing the features of the original method to ours on the Food101 dataset with 20% of training data.

REFERENCES

1. P. Khosla, P. Teterwak, C. Wang, *et al.*, “Supervised contrastive learning,” in *Advances in Neural Information Processing Systems,* vol. 33 H. Larochelle, M. Ranzato, R. Hadsell, *et al.*, eds. (Curran Associates, Inc., 2020), pp. 18661–18673.
2. O. Vinyals, C. Blundell, T. Lillicrap, *et al.*, “Matching networks for one shot learning,” in *Advances in Neural Information Processing Systems,* (2016), pp. 3630–3638.
3. T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A simple framework for contrastive learning of visual representations,” in *Proceedings of the 37th International Conference on Machine Learning,* (JMLR.org, 2020), ICML’20.
4. L. van der Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res. **9**, 2579–2605 (2008).
5. L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – mining discriminative components with random forests,” in *Computer Vision – ECCV 2014,* D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds. (Springer International Publishing, Cham, 2014), pp. 446–461.