Contrastive Learning for Fine-grained Visual Recognition

Rail Chamidullin[1]

1 Introduction

Contrastive learning is a type of representation learning which retains a representation by comparing the input samples, e.g., images, video, text, and sound. Having good representation can be beneficial for the interpretability of Deep Neural Networks (DNNs) and for some downstream tasks like open-set recognition. Contrastive learning compares positive pairs of similar inputs and negative pairs of dissimilar inputs. The key component is the contrastive loss which measures the similarity between feature vectors and enforces minimization and maximization of the similarity between positive and negative pairs. Modern contrastive learning methods are often applied in self-supervised settings, while discriminative cross-entropy learning is widely used in supervised settings. In this work, we employ supervised contrastive learning to fine-tune DNNs for fine-grained recognition.

2 Methodology

The state-of-the-art contrastive learning methods often rely only on the loss function and regularization methods like data augmentation without requiring specialized model architectures. A common approach is to use NCE-based loss motivated by standard cross-entropy-based learning. **SimCLR** (Chen et al. (2020)) is a self-supervised method that learns representations by maximizing similarity between differently augmented views of the same data example. The authors proposed a contrastive loss function using normalized-temperature cross-entropy. The pre-training method benefits from strong data augmentations, larger batch sizes, and more training steps than supervised learning. **SupCon** (Khosla et al. (2020)) learns representations similarly to SimCLR except the loss function leverages ground-truth labels making it a supervised method. To verify the theoretic assumptions about benefits of contrastive learning, we test the SupCon method over the DF20 Mini dataset with 182 visually similar species (see Figure 1) originating from 6 genera. We train various transformer-based architectures (ViT, SwinT) using a newly proposed loss function – **HCL** – that includes 2 levels of labels (species and genus) and compares additional "*soft positive pairs*" – samples from different species but the same genus.

Figure 1: Examples of inter-class similarities for selected species of two taxonomically distinct fungi genera – *Russula* and *Amanita*.

[1] PhD. student, University of West Bohemia, FAV, Department of Cybernetics, e-mail: chamirai@kky.zcu.cz

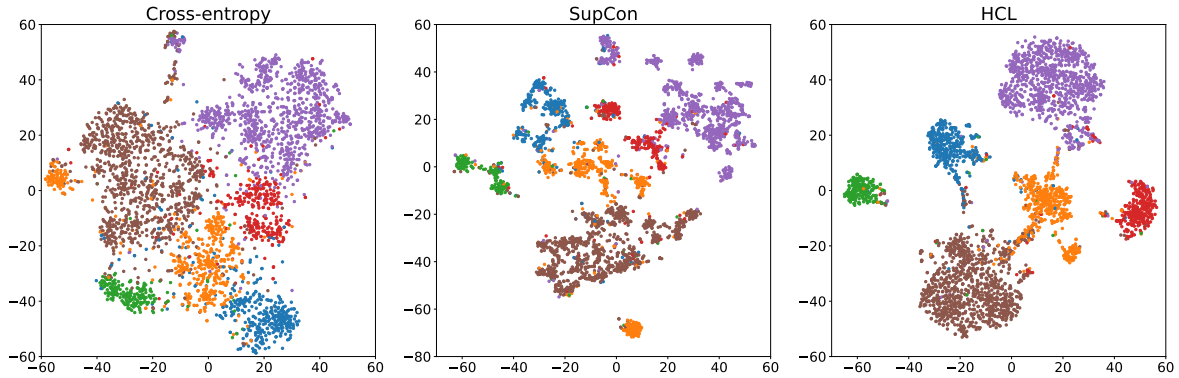

Figure 2: Feature separability on genus level achieved by different approaches. Colorized t-SNE visualization: Agaricus, Amanita, Boletus, Clitocybe, Mycena, Russula.

3 Results

The qualitative and quantitative evaluation shows significant improvement in separation for SupCon and HCL compared to a baseline cross-entropy loss. We measure the quality of representations as the number of outliers (i.e., wrong nearest centroid classifications) on a genus level. The baseline cross-entropy loss has 725 outliers (roughly 20% samples), while the SupCon and HCL losses decrease the number of outliers by 302 (41.7%) and 524 (72.3%), respectively. Changes in representations for different supervised learning losses: (i) standard cross-entropy loss; (ii) SupCon loss with species labels; and (iii) HCL with species and genus labels are visualized in Figure 2. For a detailed evaluation of individual samples, we developed an interactive t-SNE visualization that allows to view and compare input images.

4 Conclusion

This work verified the suitability and benefits of contrastive learning for fine-grained recognition. The HCL representation shown in Figure 2 has visually improved in terms of separability, and the number of outliers has decreased by up to 72.3% compared to the standard cross-entropy learning. This finding opens a promising research direction of contrastive learning for fine-grained categorization.

Acknowledgement

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The work has been supported by the grant of the University of West Bohemia, project No. SGS-2022-017.

References

Khosla, Prannay and Teterwak, Piotr and Wang, Chen and Sarna, Aaron and Tian, Yonglong and Isola, Phillip and Maschinot, Aaron and Liu, Ce and Krishnan, Dilip (2020). *Supervised Contrastive Learning*. Advances in Neural Information Processing Systems, Volume 33, pp. 18661–18673.

Chen, Ting and Kornblith, Simon and Norouzi, Mohammad and Hinton, Geoffrey (2020) *A Simple Framework for Contrastive Learning of Visual Representations*. Proceedings of the 37th International Conference on Machine Learning (ICML'20).