

FEATURE IDENTIFICATION FOR HIERARCHICAL CONTRASTIVE LEARNING

Julius Ott^{1,3}, Nastassia Vysotskaya^{2,3}, Huawei Sun^{1,3}, Lorenzo Servadei¹, Robert Wille¹

¹Technical University Munich, ²Friedrich-Alexander-Universität Erlangen, Nürnberg

³Infineon Technologies AG, Neubiberg

ABSTRACT

Hierarchical classification is a crucial task in many applications, where objects are organized into multiple levels of categories. However, conventional classification approaches often neglect inherent inter-class relationships at different hierarchy levels, thus missing important supervisory signals. Thus, we propose two novel hierarchical contrastive learning (HMLC) methods. The first, leverages a Gaussian Mixture Model (G-HMLC) and the second uses an attention mechanism to capture hierarchy-specific features (A-HMLC), imitating human processing. Our approach explicitly models inter-class relationships and imbalanced class distribution at higher hierarchy levels, enabling fine-grained clustering across all hierarchy levels. On the competitive CIFAR100 and ModelNet40 datasets, our method achieves state-of-the-art performance in linear evaluation, outperforming existing hierarchical contrastive learning methods by 2 percentage points in terms of accuracy. The effectiveness of our approach is backed by both quantitative and qualitative results, highlighting its potential for applications in computer vision and beyond.

Index Terms— Hierarchical contrastive learning, supervised learning

1. INTRODUCTION

With the rise of AI solutions in our daily life, classification remains a prominent application across various domains, such as vision [1], natural language processing [2], or discriminating generated images [3]. In the domain of machine learning-based classification, the conventional approach to learning has been to organize classes into a flat list. However, in real-world applications, hierarchical multi-labeling occurs naturally and frequently, as exemplified by its presence in biological classification (see Fig. 1), e-commerce product categorization, and retail spaces. The hierarchical representation serves to efficiently capture relationships between different classes, yet this valuable information is often underutilized in learning tasks. In representation learning frameworks, a single embedding function must generalize to unseen downstream tasks and data. Thus, this embedding function must represent the data concisely and accurately, includ-

ing the preservation of the hierarchical representation in the embedding space.

In recent years, several unsupervised [4–6] and supervised metric learning [7–9] frameworks have been proposed that rely on minimizing the distance between representations of positive pairs and maximizing the distance between negative pairs. However, these approaches frequently prove inadequate in supporting multi-label learning and leveraging information about the inter-label relationships. In this context, hierarchical multi-label contrastive learning (HMLC) methods have been proposed [10, 11]. HMLC methods augment multiple labels to single objects and impose constraints on the hierarchy of these class levels. To achieve hierarchical clustering, it is crucial to preserve the hierarchical structure of the labels in the embedding vector, where each level of the hierarchy is represented by a subset of features. The embedding vector encodes a hierarchical representation of the categories, with each level implicitly defined by a subset of features.

To enforce this hierarchical coarse-to-fine clustering, HCSC [10] learns unsupervised representations based on the highest hierarchy level. These embeddings are then clustered to select semantically unrelated negative samples. HMCE [11] is a successor that leverages supervised contrastive loss (SupCon) [7] for each level of the hierarchy, operating on one vector for all hierarchies. In turn, the fidelity of the subcategories varies significantly, with some categories having a rich hierarchical structure (e.g., "dog" can be divided into sub-categories like "golden retriever" and "poodle") while others are flatter (e.g., "wolf" has no sub-categories). This variability in fidelity highlights the need for a mechanism to identify the relevant features for each level of the hierarchy, rather than relying on a one-for-all approach. Unlike previous approaches that apply contrastive learning uniformly across all levels of the hierarchy, our method recognizes the importance of identifying relevant features for each level, and proposes a feature identification mechanism to capture nuanced relationships between categories and sub-categories. In conclusion, our proposed method represents a significant advancement in hierarchical clustering, offering a robust and flexible framework for identifying sub-level features and achieving superior cluster performance and linear evaluation. Its ability to handle disbalanced classes and

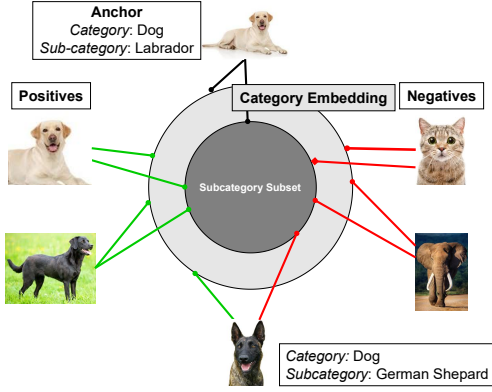


Fig. 1: Hierarchical multi-label contrastive learning (HMLC) setup. The dogs are positive pairs in the first level but negative pairs in the second level. Whereas the cat and elephant are negatives on both levels.

complex hierarchical structures makes it a promising solution for a wide range of applications. Further, its superiority over state-of-the-art approaches is clearly demonstrated by our experimental results.

2. RELATED WORK

This section first provides an overview of contrastive learning methods. Then, hierarchical multi-label contrastive learning applications and corresponding losses are presented.

2.1. Contrastive Learning

Contrastive learning has received a lot of attention as a pre-training method for self-supervised learning for images [4, 12], text [13], audio [14], and sequential data [15]. It can be summarized as *learning by comparison*. For this purpose, a data sample is chosen as an anchor. A positive sample is then of the same class, while a negative sample is from a different, “contrasting” class. The learning objective is minimizing the distance between anchor and positive sample, and maximizing it between anchor and negative sample. The key ingredients for improvements in contrastive methods are augmentations [4, 16] and batch size. Augmentations should significantly change the visual appearance while preserving semantic information. Large batch sizes are typically used, but memory issues have led researchers to investigate memory banks to access previous representations [17].

Supervised contrastive learning addresses the challenge of negative mining by using label information. The SupCon loss [7] is a generalized approach that uses multiple positives and negatives. Given a neural network encoder $E_\theta(x)$, the SupCon loss minimizes the distance between embeddings of the current sample z_i and those with the same label z_p and maximizes the distance to negative embeddings z_a , given a

temperature τ .

$$\mathcal{L}^{SupCon} = \sum_{i \in \mathcal{I}} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \quad (1)$$

However, real-world scenarios often have multiple objects per image, making a single label per image suboptimal.

2.2. Hierarchical Multi-Label Contrastive Learning

Hierarchical Multi-Label Contrastive Learning (HMLC) focuses on multiple labels per object in a hierarchical order. In a hierarchical dataset, each object has multiple labels y_i^h where $h \in 1, \dots, H$ defines the hierarchical order. This additional information enriches the label space and defines inter-label correlations.

Previous work introduced a hierarchical cost function as a regularizer to enforce hierarchical ordering [18]. The HMC loss [11] leverages the SupCon loss in the hierarchical framework, defining positive and negative pairs for each level in the hierarchy, denoted as z_p^h and z_a^h respectively. The HMC loss is defined as:

$$\mathcal{L}^{HMC} = \sum_{h=H}^1 \frac{1}{|H|} \sum_{i \in \mathcal{I}} \frac{-\lambda_h}{P^h(i)} \sum_{p \in P^h(i)} L^{pair}(i, h) \quad (2)$$

$$\mathcal{L}^{pair}(i, h) = \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p^h / \tau)}{\sum_{a \in A(i, h)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a^h / \tau)}, \quad (3)$$

where $\lambda_h = f(h)$ controls the penalty for each level in the hierarchy. The HMC loss is extended to the Hierarchical Multi-label Contrastive Enforcing (HMCE) loss, which optimizes solely with respect to the largest loss across all hierarchies. The HMCE loss is defined as:

$$\mathcal{L}^{HMCE} = \frac{1}{|H|} \sum_{h=H}^1 \sum_{i \in \mathcal{I}} \frac{-\lambda_h}{P^h(i)} \sum_{p \in P^h(i)} \max \{ \mathcal{L}^{pair}(i, h), \mathcal{L}_{max}^{pair}(i, h-1) \} \quad (4)$$

3. METHODOLOGY

In this paper, we propose feature-based hierarchical multi-label contrastive learning methods. Intuitively, it is reasonable to perform contrastive learning only on that part of the embedding vector that is relevant for a certain hierarchy level. To this end, we propose two ways to identify the relevant features, which are detailed in Figure 2. First, we apply a Gaussian Mixture Model (GMM), which masks relevant features in the embedding before contrastive learning (G-HMLC) is performed. To this end, two embedding vectors are generated: the anchor image and the same image with random augmentations. These vectors define a two-dimensional plane, where a two-dimensional GMM is fitted with prior knowledge about the number of hierarchy levels. The GMM then predicts the

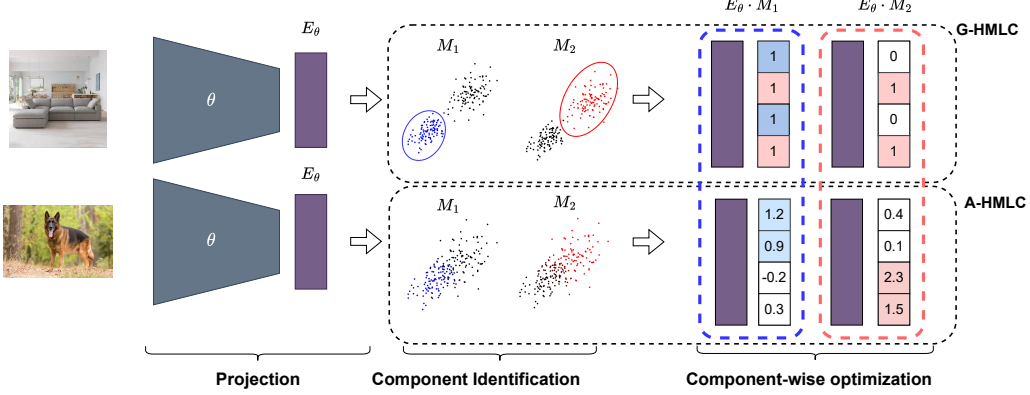


Fig. 2: Illustration of the G-HMLC and A-HMLC architectures. The projection head E_θ maps the images to embedding vectors. In G-HMLC, a GMM is fitted on this embedding vector to generate a mask for each hierarchy level. This hard masking is suitable for unrelated lower hierarchy classes (furniture-couch $\neq$ furniture-table). For shared features along the hierarchy tree (e.g. dogs), A-HMLC computes soft attention scores. The hierarchical binary or soft masks, both denoted as M_i for readability, are then multiplied by the feature vector.

masks M_h for each hierarchy level. These masks are incorporated into Eq. 4, where we replace $\mathbf{z}_i$ with $\mathbf{z}_i[M_h]$, $\mathbf{z}_a$ with $\mathbf{z}_a[M_h]$ and $\mathbf{z}_p^h$ with $\mathbf{z}_p^h[M_h]$ and define it as $\mathcal{L}^{pair}(i, h)[M_h]$. Then the SupCon loss is applied to the masked feature vectors for each level. The GMM is applied during training for each image in the batch individually, and is initialized with the previously identified mean values. This setup optimizes the convergence speed, which further reduces the number of iterations. Intuitively, this hard masking performs well when the lower hierarchy samples are clearly distinguishable, for instance, in 3, or when we aim to distinguish different furniture, it is clear that a wardrobe is significantly different from a chair.

However, some data points share features across the hierarchy tree. To address this, we propose the second approach: attention-based multi-label hierarchical contrastive loss (A-HMLC), using soft masking via multiplicative attention maps. Following the self-attention implementation of [19], we define K , Q as single-layer neural networks, which take the embedding vector as input. The attention weights are defined as:

$$attention = softmax\left(\frac{Q \cdot K}{\sqrt{d}}\right), \quad (5)$$

where d is the dimension of the embedding vector. The attention weights are multiplied by the embedding vector before the contrastive loss is applied. In addition, we employ one attention head per hierarchy level, such that each head learns the relevant features for the specific level. In both approaches, a linear scaling was employed compared to the exponential scaling in the HMCE loss.

4. EXPERIMENTS

The subsequent experiments provide a comprehensive evaluation of the Gaussian (G-HMLC) and attention-based (A-



Fig. 3: Examples of the hierarchical MNIST dataset. The central digit refers to the class, and the image size is scaled from 32×32 to 192×192 pixels. The subsidiary digit around denotes the category and is placed randomly with a size of 32×32 .

HMLC) feature extraction methods, assessing their performance through both qualitative and quantitative analyses.

4.1. Qualitative Evaluation

The experiments under consideration encompass the state-of-the-art hierarchical contrastive HMCE loss with our proposed G-HMLC loss. In this experiment, the objective is to evaluate the clustering quality of the hierarchical contrastive losses for each hierarchy level. To this end, a hierarchical MNIST dataset is created, as shown in Figure 3. Note that the category and sub-category are distinct, making the G-HMLC the appropriate choice here. In Figure 4 the first and second t-SNE [20] components of the test dataset embeddings are presented. The HMCE loss demonstrates effective separation at the first level but exhibits a lack of distinction at the second level. This observation underscores the significance of the ordering of the hierarchy in the HMCE loss formulation, and a noteworthy performance is anticipated when the class is at the first level and the categories represent the higher-order categories. The proposed G-HMLC loss attains a remarkable first- and second-level separation. At the same time, 3% more variance in the embeddings are expressed by the G-HMLC loss, underlining the more effective dimensionality reduction.

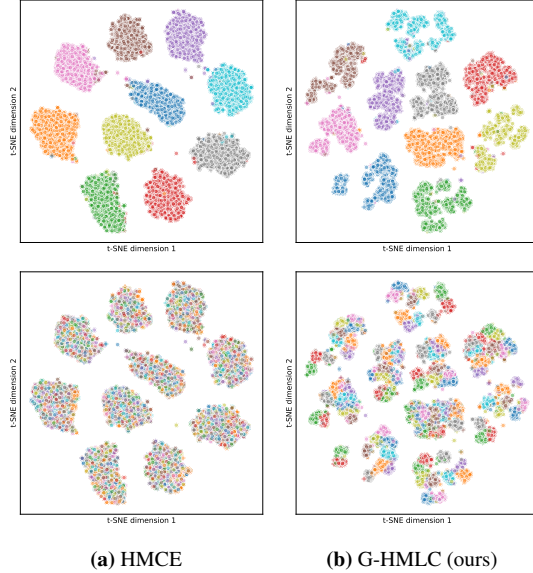


Fig. 4: First and second *t*-SNE components of the hierarchical MNIST test embeddings. The HMCE loss (left column) separates the first level (top row) but misses a clear separation for the second level (bottom row). The proposed G-HMLC loss (right column) separates both levels.

4.2. Quantitative Evaluation

The experiments under consideration encompass the hierarchical contrastive losses. As a baseline, we utilize the single-label contrastive SupCon loss and non-contrastive cross-entropy loss. To assess its efficacy for more complex tasks and diverse data modalities, we examined the balanced CIFAR100 [21] image dataset and the unbalanced ModelNet40 [22] point-cloud dataset. The CIFAR100 dataset comprises 100 classes, each originating from 20 distinct categories, where each category has 5 sub-classes. On the contrary, the ModelNet40 dataset is grouped into 6 categories with the number of sub-categories ranging from 2 to 26. We conduct a ResNet50 [23] as an encoder for CIFAR100 and a PointNet++ encoder [24] for ModelNet40. The test accuracy values presented in Table 1 have been obtained in accordance with the linear evaluation scheme outlined in [4]. The proposed A-HMLC and G-HMLC losses improve the performance on the balanced CIFAR100 dataset. As assumed in the introduction, the attention-based A-HMLC loss further outperforms the G-HMLC loss since many classes share features across hierarchy levels. In addition, all hierarchical contrastive variants outperform the single-label losses. In contrast, the HMCE loss falls short against the single-label losses using the unbalanced ModelNet40 dataset. This highlights the aforementioned issue on the fidelity of the sub-classes, which is not addressed by the HMCE loss. Since the proposed methods update each feature independently, they are more robust to the unbalanced data.

Method	ModelNet40	CIFAR100
Cross Entropy	90.5	75.3
SupCon	91.2	75.26
HMCE	90.2	75.95
G-HMLC (ours)	91.5	76.13
A-HMLC (ours)	92.42	76.19

Table 1: Test accuracy after linear evaluation on ModelNet40 and CIFAR100 datasets.

Method	ModelNet40	CIFAR100
HMCE category first	85.4	73.21
HMCE class first	90.2	75.95
G-HMLC category first	91.05	74.36
G-HMLC class first	91.5	76.13
A-HMLC category first	92.38	75.69
A-HMLC class first	92.42	76.19

Table 2: Ablation on the hierarchical order.

4.3. Hierarchical Ordering

To assess the impact of hierarchy ordering, an ablation study was conducted, where the ordering of the labels is reversed, which impacts the algorithmic behavior due to hierarchical scaling. In the presence of a perfect hierarchical clustering, the order ought to be irrelevant. The numerical evaluations, presented in Table 2, demonstrate that the G-HMLC and A-HMLC exhibit enhanced robustness compared to the HMCE loss. This quantitative finding is corroborated by the clustering quality exhibited in Figure 4. Consequently, it can be deduced that feature-based HMC losses possess the capacity to learn fine-grained hierarchical clusters that are resilient to label order variations.

5. CONCLUSION

This paper proposes feature-based losses for hierarchical contrastive learning. The proposed variants, G-HMLC via clustering for small datasets and the end-to-end A-HMLC, operate on the relevant subsets of the embedding space for the respective hierarchical features. This enables fine-grained clustering on all hierarchy levels. As shown in qualitative and quantitative experiments, the superior clustering performance is backed by a 2% improved accuracy on the linear evaluation for balanced and unbalanced hierarchical datasets. A subsequent study on the order of the hierarchy levels demonstrated that feature-based losses effectively utilize all supervisory labels. In this study, we assume that the number of hierarchy levels is known for the employed clustering method. In future work, we will explore the identification of unknown hierarchy levels using a GMM with a Dirichlet process to identify an arbitrary number of clusters in the embedding.

6. REFERENCES

- [1] Peiyuan Jiang et al., “A review of yolo algorithm developments,” *Procedia computer science*, vol. 199, pp. 1066–1073, 2022.
- [2] Alper Kursat Uysal and Serkan Gunal, “The impact of preprocessing on text classification,” *Information processing & management*, vol. 50, no. 1, pp. 104–112, 2014.
- [3] Jonas Adler and Sebastian Lunz, “Banach wasserstein gan,” *Advances in neural information processing systems*, vol. 31, 2018.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, “A simple framework for contrastive learning of visual representations,” in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.
- [5] Mathilde Caron et al., “Emerging properties in self-supervised vision transformers,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
- [6] Mathilde Caron et al., “Unsupervised learning of visual features by contrasting cluster assignments,” *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.
- [7] Prannay Khosla et al., “Supervised contrastive learning,” *Advances in neural information processing systems*, vol. 33, pp. 18661–18673, 2020.
- [8] Lorenzo Servadei et al., “Label-aware ranked loss for robust people counting using automotive in-cabin radar,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 3883–3887.
- [9] Gal Chechik, Varun Sharma, Uri Shalit, and Samy Bengio, “Large scale online learning of image similarity through ranking,” *Journal of Machine Learning Research*, vol. 11, no. 36, pp. 1109–1135, 2010.
- [10] Yuanfan Guo et al., “Hcsc: Hierarchical contrastive selective coding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9706–9715.
- [11] Shu Zhang et al., “Use all the labels: A hierarchical multi-label contrastive learning framework,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16660–16669.
- [12] Sumit Chopra, Raia Hadsell, and Yann LeCun, “Learning a similarity metric discriminatively, with application to face verification,” in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*. IEEE, 2005, vol. 1, pp. 539–546.
- [13] Alec Radford et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
- [14] Alexei Baevski et al., “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.
- [15] Pierre Sermanet et al., “Time-contrastive networks: Self-supervised learning from video,” in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.
- [16] Ekin D Cubuk et al., “Autoaugment: Learning augmentation policies from data,” in *2019 IEEE computer society conference on computer vision and pattern recognition (CVPR’19)*, 2019.
- [17] Kaiming He et al., “Momentum contrast for unsupervised visual representation learning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.
- [18] Loic Landrieu and Vivien Sainte Fare Garnot, “Leveraging class hierarchies with metric-guided prototype learning,” in *British Machine Vision Conference (BMVC)*, 2021.
- [19] Ashish Vaswani et al., “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [20] Laurens van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.
- [21] A. Krizhevsky, “Learning Multiple Layers of Features from Tiny Images,” Technical report, Univ. Toronto, 2009.
- [22] Zhirong Wu et al., “3d shapenets: A deep representation for volumetric shapes,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1912–1920.
- [23] Kaiming He et al., “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [24] Charles Ruizhongtai Qi et al., “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.