Learning Hierarchy Aware Features for Reducing Mistake Severity

Ashima Garg, Depanshu Sani, and Saket Anand

Indraprastha Institute of Information Technology, Delhi, India
{ashimag, depanshus, anands}@iiitd.ac.in

Abstract. Label hierarchies are often available apriori as part of biological taxonomy or language datasets WordNet. Several works exploit these to learn hierarchy aware features in order to improve the classifier to make semantically meaningful mistakes while maintaining or reducing the overall error. In this paper, we propose a novel approach for learning *Hierarchy Aware Features (HAF)* that leverages classifiers at each level of the hierarchy that are constrained to generate predictions consistent with the label hierarchy. The classifiers are trained by minimizing a Jensen-Shannon Divergence with target soft labels obtained from the fine-grained classifiers. Additionally, we employ a simple geometric loss that constrains the feature space geometry to capture the semantic structure of the label space. HAF is a *training time* approach that improves the mistakes while maintaining top-1 error, thereby, addressing the problem of cross-entropy loss that treats all mistakes as equal. We evaluate HAF on three hierarchical datasets and achieve state-of-the-art results on the iNaturalist-19 and CIFAR-100 datasets. The source code is available at `https://github.com/07Agarg/HAF`

1 Introduction

Conventional classifiers trained with the cross-entropy loss treat all misclassifications equally. However, certain categories may be more semantically related to each other than to other categories, implying that some classification mistakes may be more *severe* than others. For instance, an autonomous vehicle confusing a `car` for a `truck` is not as severe as mistaking a `pedestrian` for `road`, where the latter mistake could lead to a catastrophe. Similarly, falsely identifying a `pine tree` with an `oak tree` is less severe than identifying it as a `rose`. Classifiers trained to make mistakes with lower severity could benefit and are often critical in many real-world applications.

The severity of a mistake is typically defined based on some notion of semantic similarity between class labels. For example, a taxonomic hierarchy tree defined over the class labels can express specific semantic relationships between classes through its tree structure, thus enabling an ordering of classes. This ordering was obtained using the lowest common ancestor (LCA) measure in [4, 14]. These hierarchies are often readily available in the class label space as part of language

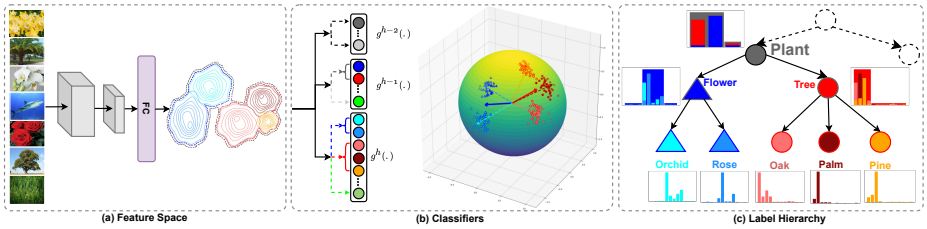


Fig. 1: **Overview of HAF**. We propose a probabilistic approach using to learn *hierarchy-aware features* that respect the label hierarchy in the feature space and thereby make semantically meaningful mistakes. We train separate classifiers, with a shared feature space, for each level of the label hierarchy. We model the relationship between the fine-grained classes and their respective coarser classes using the label hierarchy and impose consistency constraints on the probability distributions. We further impose simple geometric constraints on the weight vectors of classifiers from different levels to align the weight vectors of fine-grained classes with their corresponding weight vectors of coarser-classes.

datasets like WordNet [19] or from biological taxonomies, e.g., the one used with the iNaturalist-19 dataset [25].

Bertinetto et al. [4] proposed approaches to reduce the severity of mistakes by employing hierarchy-sensitive adaptations of the cross-entropy loss. They reported reduction in the mistake severity based on the average hierarchical distance of top-k predictions at the cost of an increased top-1 error, with the trade-off being controlled by a hyperparameter. A more desirable solution would be the one that reduces the severity of mistakes while maintaining or reducing the overall top-1 error. Karthik et al. [14] highlighted this trade-off and pointed out that the classical approach of Conditional Risk Minimization (CRM) could reduce the mistake severity without a significant change in the top-1 error. Moreover, CRM is a test-time intervention that applies post-hoc corrections on the class likelihoods using the LCA measure between classes. Despite its simplicity, the CRM approach is versatile and its effectiveness is remarkable. In Sec. 4, we show that CRM, when combined with other approaches, almost always improves the mistake severity, without a significant impact on the top-1 error.

While CRM improves the quality of prediction errors, being a test-time approach, it does not affect the model. Consequently, the learned representations are inherently inadequate because the cross-entropy loss function ignores all semantic structure in the label space and treats each class independently. To overcome this limitation, the hierarchical cross-entropy (HXE) loss was proposed in [4], which essentially amounts to a weighted combination of the cross-entropy loss applied at different levels of the hierarchy[1]. Chang et al. [7] pointed out that training with a coarse class cross-entropy loss deteriorates the accuracy at fine-grained levels. This is likely the reason why both variants proposed in [4],

[1] See the supplementary material for a derivation

HXE and the soft-labels loss, result in a trade-off between top-1 error and the severity of mistakes. Chang et al. [7] mitigate this trade-off by disentangling the coarse and the fine-grained features by explicitly partitioning the feature space. This disentangling approach proved to be successful for small hierarchies, however, the feature vector partitioning limits its scalability to larger hierarchies. We argue that for addressing the problem of mistake severity, while maintaining the top-1 error, it is important to learn a feature space that captures the structure available in the label space. To this end, we propose learning a *hierarchy-aware feature* (HAF) space that is explicitly trained to inherit the hierarchical structure of the labels.

We observe that a hierarchy-aware feature space should enable classification at *all* levels of the hierarchy, and simultaneously lead to a lower mistake severity at the finest level. The label hierarchy structure constrains the coarse-level class labels to be a composition of *disjoint* sets of its sub-classes in the hierarchy. We exploit two key properties of the classifiers acting on the feature space to help inherit this compositional structure from the label space.

First, we train a classifier using the fine-grained cross-entropy and use its predictions to obtain target soft labels (Fig. 1) for training the coarse-level auxiliary classifiers. The coarse-level classifiers minimize the Jensen-Shannon divergence (JSD) between their predictions and the target soft labels. This loss avoids the use of hard labels at coarser levels and thus serves as a consistency regularization for the fine-grained classifier, which in turn leads to improved mistake severity without compromising the top-1 error. We take this approach to avoid the pitfall highlighted in [7], which states that fine-grained features can lead to better coarse-grained predictions, however, explicitly using cross-entropy loss for coarse-level classifiers leads to feature spaces that worsens the performance at a finer granularity. Second, we impose geometric consistency constraints on the classifier weight vectors that align sub-classes belonging to the same super-class (Fig. 1(b)). The resulting loss promotes a feature space (Fig. 1(a)) that respects the semantic hierarchy of the label space (Fig. 1(c)). We present further details of the loss terms in Sec. 3. We summarize our contributions below.

- We introduce a novel approach for learning a *hierarchy-aware feature* (HAF) space by inheriting the structure of the label space. We design the loss functions that impose probabilistic and geometric constraints between coarse and fine level classifiers.
- We empirically demonstrate that HAF scales well with large label hierarchies and reduces mistake severity while maintaining the top-1 fine-grained error.

2 Related Work

Several works exploit the hierarchical taxonomy of the data for image classification for visual [4, 14, 7] and text [18] data, multi-label classification tasks [27], image retrieval [2, 29], object recognition [22], and recently to improve semi-supervised approaches [12, 24]. We discuss some of the important works that are closely related with our objective.

Label-embedding methods. These methods model the class relationships using soft-embeddings. DeViSE [10] maximizes the cosine similarity between the embeddings of an image extracted from a pretrained visual model and the embeddings of label obtained using pretrained word2vec model on Wikipedia. Liu et al. [17] exploit hyperbolic geometry to learn the hierarchical representations. Similar to DeViSE [10], they minimize the Poincaré distance between the Poincaré label embeddings [20] and the image features embeddings. Barz & Denzler [2] map the embeddings onto a unit hypersphere and use LCA to encode the hierarchical distances. Bengio et al. [3] impose the structure over the classes and fastens learning to embed in low dimensional space to model semantic relationships between classes. Bertinetto et al. [4] proposed Soft-labels that uses the soft-targets encoded with inter-class semantic information based on LCA.

Hierarchical-architecture based methods. Wu et al. [28] jointly optimize multi-task loss function wherein cross-entropy loss is applied at each hierarchical level. Recently, Chang et al. [7] established that jointly optimizing fine-grained with coarse-grained recognition in vanilla framework deteriorates performance on fine-grained classification. The authors proposed architecture for multi-granularity classification with independent level-specific classifiers. Redmon et al. [22] proposed a probabilistic model, YOLOv2, for object detection and classification, where softmax is applied at every coarse-category level to address the mutual exclusion of all the classes in conventional softmax classifier.

Hierarchical-loss based methods. Bertinetto et al. [4] proposed another approaches - hierarchical cross-entropy (HXE). HXE is a probabilistic approach that optimizes a loss function based on conditional probabilities, where predictions for a particular class is conditioned on the parent-class probabilities. Brust & Denzler et al. [6] proposed a conditional probability classifier for DAGs. Bilal et al. [5] proposed hierarchical-aware convolutional neural networks by adding branches to the intermediate network pipeline. In [16], authors use prototypical network which uses softmax over distances between the features to the class prototypes, along with a regularization term that encourages the class prototypes to follow the relationship in label hierarchy. Our work is in line with this body of research. We study a different probabilistic model and propose a loss function based on that model. In HAF, we explicitly define class prototypes at every level and take a different approach for arrangement of these prototype vectors.

Cost based methods. Another line of research is based on assigning different costs depending on the types of misclassification [1]. Deng et al. [8] proposed to use *mean classification cost* to make hierarchy-aware predictions by penalizing the mistakes based on the hierarchies. [9, 26] used semantic hierarchy to design cost matrix optimizing accuracy-specificity trade-offs between the level of abstraction of the selected class while selecting the best in specificity. These methods include both internal and leaf nodes in the cost matrix. While Karthik et al. [14] study conditional risk minimization (CRM) on similar lines to [8], an inference-time approach that weighs the predictions based on the cost matrix defined using LCA distances among the leaf nodes. HAF also fits in this frame-

work. However, unlike CRM [14], HAF is a training-time approach to learn feature embeddings such that they are hierarchically meaningful.

3 HAF: Proposed Approach

Consider a label hierarchy tree with $H+1$ levels, where the root is at level-0, and $h \in [1, \ldots, H]$ denote the hierarchical level with $h=1$ and $h=H$ the coarsest and finest levels respectively. We ignore the root node for our purposes as it denotes the universal super-set containing all classes. Let $\mathcal{X} = \{\mathbf{x}_i, y_i^h | i = 1, \ldots, N\}$ be the set of N images and their respective ground-truth labels at level h. We denote the common feature extractor $f_\phi(\cdot)$, which is implemented using some backbone neural network and is parameterized by ϕ. As illustrated in Fig. 1, we use classifiers at each level of the hierarchy in training HAF and denote the level-h classifier as $g^h(\cdot)$ parameterized by the weight matrix $\mathbf{W}^h$. The resulting prediction probabilities are denoted by $p^h(\widehat{y}_i^h | \mathbf{x}_i; \mathbf{W}^h) = g^h(f_\phi(\mathbf{x}_i))$, where $\widehat{y}_i^h$ is the label predicted for $\mathbf{x}_i$ by $g^h(\cdot)$ and can take class labels from the set of classes at level-h as $\mathcal{C}^h = \left\{\bigcup_{i=1}^{|A|} A_i, \bigcup_{i=1}^{|B|} B_i, \bigcup_{i=1}^{|C|} C_i, \ldots\right\}$, where we define the set of classes at level-$(h-1)$ as $\mathcal{C}^{h-1} = \{A, B, C, \ldots\}$. With a slight abuse of notation, here we use A to denote a super-class label at level-$(h-1)$ and the set of its sub-classes $\{A_1, A_2, \ldots\}$ at level-h.

3.1 Fine Grained Cross-entropy ($L_{CE_{fine}}$)

We use the ground truth labels only at the finest level of the hierarchy and apply the cross-entropy loss to train the level-H classifier, i.e., $g^H(\cdot)$. The fine-grained cross-entropy loss for a sample is given by

$$L_{CE_{fine}} = -\sum_{c \in \mathcal{C}^H} \mathbf{1}\left[y_i^H = c\right] \log\left(p^H(\widehat{y}_i^H = c | \mathbf{x}_i; \mathbf{W}^H)\right) \tag{1}$$

where $\mathbf{1}[\cdot]$ serves as an indicator function and takes a value of one when the argument is true, else zero.

3.2 Soft Hierarchical Consistency (L_{shc})

For making better mistakes, we want the classifiers at all levels to use the same feature space and yet make predictions consistent with the label hierarchy. While it is natural to use the cross-entropy loss for training the classifiers at all levels, as noted in [7] and observed during our initial experiments, this choice of loss compromises the fine-grained accuracy. Instead, we enforce the consistency across classifiers at different levels by using soft labels and a symmetric entropy-based loss function. We minimize the Jensen-Shannon Divergence (JSD) [11] between the predictions of a coarse classifier $g^{h-1}(\cdot)$ and the soft labels obtained from the next fine-level classifier $g^h(\cdot)$. As defined above, for a given class label $A \in \mathcal{C}^{h-1}$,

let $P[\widehat{y}^{h-1}=A|\mathbf{x}_i]$ denote the probability of the sample $\mathbf{x}_i$ belonging to the class A, which is computed as

$$P\left[\widehat{y}_i^{h-1}=A|\mathbf{x}_i\right]=\sum_{k=1}^{|A|}p^h(\widehat{y}_i^h=A_k|\mathbf{x}_i;\mathbf{W}^h) \tag{2}$$

The probabilities $P[c],\forall c\in\mathcal{C}^{h-1}$ are concatenated together to construct the probability vector $\widehat{p}^{h-1}(\widehat{y}_i^{h-1}|\mathbf{x}_i)$, which is used as the soft label for $\mathbf{x}_i$. This soft label generation process is illustrated in Fig. 2.

The JSD is minimized between the soft labels and the predictions from the classifier $g^h(\cdot)$. For convenience, we use p_i^h to refer to $p^h(\hat{y_i}^h|x_i;\mathbf{W}^h)$ and similarly $\widehat{p}_i^h$ for the corresponding soft label. The JSD based total Soft Hierarchical Consistency is computed by summing the pairwise losses across the levels

$$\mathrm{L}_{shc}=\sum_{h=1}^{H-1}JS^h\left(p_i^h||\widehat{p}_i^h\right)=\frac{1}{2}\sum_{h=1}^{H-1}(\mathrm{KL}(p_i^h||m)+\mathrm{KL}(\widehat{p}_i^h||m)) \tag{3}$$

where $m=\frac{1}{2}(p_i^h+\widehat{p}_i^h)$ and $\mathrm{KL}(\cdot||\cdot)$ refers to Kullback-Leibler divergence.

It is important to highlight the key difference between the soft labels generated above and those defined in [4]. The latter are designed using the LCA-based distance between classes, whereas our choice of soft labels can be interpreted as a *learned* label-smoothing that better regularizes the coarse-level classifiers. Yuan et al. [30] make a similar argument about label smoothing in the context of knowledge distillation. The use of a symmetric loss like in eqn. (3) further enables the classifiers at both levels to jointly drive the feature space learning. This behavior of the coarse classifiers improving the performance of the finer-level classifiers is analogous to the Reversed Knowledge Distillation (Re-KD) setting as presented in [30], where the authors showed that a student ($g^{h-1}(\cdot)$) is capable of improving the performance of the teacher ($g^h(\cdot)$).

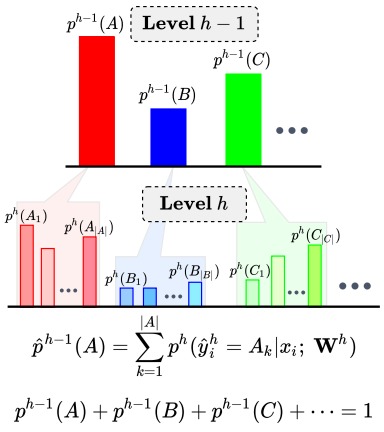


Fig. 2: Constructing the soft labels for training the coarse-level classifiers. The super-class target probability is the sum of its sub-classes' predicted probability. The colors indicate the class relationships across levels $h-1$ and h.

3.3 Margin Loss (L_m)

While L_{shc} improves the mistake severity (as we show in Sec. 5) successfully by virtue of better regularization, it does not directly encourage discrimination between coarse-level classes. Therefore, we use a pairwise margin-based loss to

promote a more discriminative feature space. We use this loss over coarser levels $h \in \mathcal{H}$ where $\mathcal{H}$ is $[k,\ H-1]$ and k ranges from $[1\ ,\ H-1]$. For a given batch of samples, we create pairs of samples that have dissimilar labels at a level h, i.e., $\mathcal{B}^h = \{(i,j)|y_i^h \neq y_j^h\}$. Then we compute the margin loss over the batch as

$$L_m = \sum_{h \in \mathcal{H}} \sum_{(i,j) \in \mathcal{B}^h} \max\left(0, m - \mathrm{JS}^h(p_i^h || p_j^h)\right) \tag{4}$$

where p_i^h is the softmax probability generated by $g^h(f_\phi(\mathbf{x}_i))$ and m is the margin. The margin loss is only applied to the coarser levels of the hierarchy, as the cross-entropy loss of (1) is sufficient for fine-grained discrimination.

3.4 Geometric Consistency (L_{gc})

HAF uses classifiers at all the levels of hierarchy. In a hierarchy-aware feature space, the weight vectors of the coarse class and its fine-grained classes should be correlated. The losses introduced in the previous subsections impose probabilistic consistency across the classifier predictions, and only indirectly affect the feature space geometry. In order to better orient the feature space to inherit the label space hierarchy, we use a geometric consistency loss. As before, let $A \in \mathcal{C}^{h-1}$ be a given super-class and its sub-classes be $A_k \in \mathcal{C}^h,\ k = 1, \ldots, |A|$. Let the weight vector corresponding to the super-class A be $\mathbf{w}_A^{h-1}$ and similarly the weight vectors corresponding to the sub-classes be $\mathbf{w}_{A_k}^h$. Note that the classifier $g^{h-1}(\cdot)$ is defined by the weight matrix $\mathbf{W}^{h-1}$, which is obtained by stacking the weight vectors $\mathbf{w}_c,\ c \in \mathcal{C}^{h-1}$. We further constrain each weight vector to be unit norm $||\mathbf{w}_c^h||_2 = 1, \forall c, h$, across all classifiers. For the super-class $A \in \mathcal{C}^{h-1}$, we define the target weight vector as $\widehat{\mathbf{w}}_A^{h-1} = \widetilde{\mathbf{w}}_A^{h-1} / ||\widetilde{\mathbf{w}}_A^{h-1}||_2$, where $\widetilde{\mathbf{w}}_A^{h-1} = \sum_{k=1}^{|A|} \mathbf{w}_{A_k}^h$. Thus, the Geometric Consistency loss to be minimized is

$$L_{gc} = \sum_{h=1}^{H-1} \sum_{c \in \mathcal{C}^h} \left(1 - \cos\left(\mathbf{w}_c^h, \widehat{\mathbf{w}}_c^h\right)\right) \tag{5}$$

where $\cos\left(\mathbf{w}_c^h, \widehat{\mathbf{w}}_c^h\right)$ refers to the cosine similarity between the weight vectors $\mathbf{w}_c^h$ and $\widehat{\mathbf{w}}_c^h$.

Finally, the total loss is given by $L_{total} = L_{CE_{fine}} + L_{shc} + L_m + L_{gc}$.

4 Experiments and Results

4.1 Experimental Setup

Datasets. We present the evaluation of HAF approach on the CIFAR-100 [15], iNaturalist-19 [25] and tieredImageNet-H [23] datasets. We follow the hierarchical taxonomy as is in [16] for CIFAR-100, and [4] for iNaturalist-19 and tieredImageNet-H. In all the three datasets, Level-0 has only one node, i.e., the

root node. Therefore, we only consider the bottom H hierarchical levels. Similar to [4], we compute the distance between any two nodes by finding the minimum distance between the node and their Lowest Common Ancestor (LCA). Table 1 summarizes the dataset statistics.

	Train	Val	Test	#Classes	#Levels
CIFAR-100	45,000	5,000	10,000	100	6
iNaturalist-19	187,385	40,121	40,737	1010	8
tieredImageNet-H	425,600	15,200	15,200	608	13

Table 1: Statistics of the datasets.

Baselines. We directly compare HAF with the baseline cross-entropy, Barz & Denzler's [2], YOLO-v2 [22], both approaches of Bertinetto et al's [4] work - soft-labels and HXE, and the recently proposed CRM-based method from [14]. We also compare with recently proposed Chang et al.'s [7] multi-task framework for classification with different granularities. For fair comparisons, we re-run all the experiments with the same codebase under the new best hyperparameter settings for all the methods and report mean and standard deviation of each experiment averaged over three-different seeds.
Evaluation Metrics. We use the same evaluation metrics as Bertinetto et al. [4]; Karthik et al. [14]. We report the following three metrics: i) top-1 error, ii) average mistakes severity, i.e., average LCA-based distance between the ground-truth and predicted class label for *only* incorrectly classified samples, and iii) average hierarchical distance @k, i.e., average distance from the LCA of ground-truth label and k most likely predictions for *all* the samples.

4.2 Training Configurations

We adopt the Wideresnet-28-2 [31] backbone for evaluation on the CIFAR-100 dataset. For the iNaturalist-19 and tieredImageNet-H datasets, we use the ImageNet pretrained ResNet-18 [13] backbone with an additional fully-connected (FC) layer of 600 hidden units. Chang et al. [7] only employ this fully connected layer for facilitating disentanglement, however, we use this additional layer as part of the backbone for consistency across all the methods. Classifiers for each hierarchical level follow this layer. We train all the models with a batch size of 256. We use a fixed margin m of 3.0 across all the datasets defined in Eq (4) and create a total of 256 dissimilar pairs from a batch of data. For CIFAR-100, we employ `RandomPadandCrop(32)` and `RandomFlip()` for augmentation. For iNaturalist-19 and tieredImagenet-H, we use `RandomHorizontalFlip()` followed by `RandomResizedCrop()` as carried out in [4].

We find the training strategy (learning rate and optimizer) of Chang et al. [7] to give optimal results on both CIFAR-100 and iNaturalist-19 datasets on the baseline cross-entropy. This training strategy with the SGD optimizer boosts the

performance of cross-entropy on iNaturalist-19 as opposed to the ones reported using Adam optimizer in [4]. We obtain the best results for CIFAR-100 and iNaturalist-19 using the SGD optimizer on all methods, except for soft-labels and HXE where Adam [21] performs the best. For the methods trained with SGD, we set different learning rates for the backbone network and the FC layer as 0.01 and 0.1 respectively, following [7]. For training with soft-labels and HXE with Adam optimizer, using a hyperparameter sweep we find that the model performs the best with learning rate as $1e-3$ and $1e-4$ for CIFAR-100 and iNaturalist-19 respectively. We train all the models on tieredImageNet-H for 120 epochs with a learning rate of $1e-5$. Unlike other datasets, we employ the Adam optimizer for tieredImageNet-H as it performed better than the SGD optimizer.

4.3 Results

Tables 2, 3 and 4 present the comparisons of our proposed technique with the baselines on CIFAR-100, iNaturalist-19, and tieredImageNet-H respectively. Karthik et al. [14] apply the CRM technique on the baseline cross-entropy. Since CRM is a test-time approach that reweighs the probability distribution of samples obtained from any trained model, it can be applied to all other approaches. Therefore, in each of the Tables 2-4, we group the results to report evaluation metrics with and without using CRM at test-time. We re-emphasize that the goal of the problem is to improve the hierarchical metrics by maintaining or improving the top-1 error. Towards this goal, in each table, we highlight the competitive methods (rows) on the top-1 error with lightgreen. Among these competitive methods, we highlight the best-performing entries for each metric with green. On CIFAR-100 (Table 2), baseline cross-entropy, Chang et al.[7] and HAF and their counterparts using CRM are competitive methods on top-1 error. However, HAF and HAF + CRM outperforms all other hierarchical metrics without compromising top-1 error. We observe similar trends on iNaturalist-19 (Table 3), where HAF , and HAF + CRM are the only competitive *training method* to cross-entropy, which maintain the top-1 error and yet improve the hierarchical metrics. On tieredImageNet-H (Table 4), baseline cross-entropy, HXE $\alpha = 0.1$, Soft-labels $\beta = 30$, and HAF are competitive for both, top-1 error and hierarchical metrics. However, HAF is the best performing method on hier dist@20.

It is worth pointing out that Chang et al.'s [7] method does not scale well with increasing number of hierarchical levels. For CIFAR-100 with six levels, the accuracy is competitive with cross-entropy, however, with both iNat and tieredImageNet-H, which have 8 and 13 levels, the top-1 error worsens. This is not unexpected as the feature vector is divided based on number of levels. While increasing the feature space may be a reasonable solution to maintain performance, it may not be straightforward to decide the feature vector size for each level, especially for hierarchies that may be skewed. On the contrary, HAF is independent of the number of hierarchical levels used despite using hierarchical classifiers at each level. We also note that the CRM approach fails to improve Soft-labels β=4. This is perhaps because the label distribution is very flat for

Method	Top-1 Error(↓)	Mistakes severity(↓)	Hier dist@1(↓)	Hier dist@5(↓)	Hier dist@20(↓)
			Without CRM		
Cross-Entropy	22.27 ± 0.001	2.35 ± 0.024	0.52 ± 0.003	2.24 ± 0.007	3.17 ± 0.007
Barz & Denzler	31.69 ± 0.004	2.36 ± 0.025	0.75 ± 0.012	1.25 ± 0.364	2.49 ± 0.004
YOLO-v2 [22]	32.03 ± 0.006	3.72 ± 0.022	1.19 ± 0.019	2.85 ± 0.010	3.39 ± 0.0109
HXE α=0.1 [4]	28.41 ± 0.003	2.43 ± 0.004	0.69 ± 0.008	2.08 ± 0.008	3.02 ± 0.012
HXE α=0.6 [4]	30.42 ± 0.003	2.29 ± 0.008	0.7 ± 0.008	1.76 ± 0.007	2.79 ± 0.008
Soft-labels $\beta = 30$ [4]	26.99 ± 0.003	2.38 ± 0.004	0.64 ± 0.008	1.39 ± 0.027	2.79 ± 0.005
Soft-labels $\beta = 4$ [4]	32.15 ± 0.008	2.21 ± 0.037	0.71 ± 0.024	1.23 ± 0.018	2.23 ± 0.008
Chang et al. [7]	21.94 ± 0.002	2.32 ± 0.005	0.51 ± 0.005	2.06 ± 0.018	3.08 ± 0.007
HAF	22.27 ± 0.001	2.24 ± 0.014	0.50 ± 0.003	1.41 ± 0.007	2.64 ± 0.002
			With CRM		
Cross-Entropy [14]	22.23 ± 0.001	2.31 ± 0.033	0.51 ± 0.006	1.11 ± 0.006	2.18 ± 0.002
YOLO-v2	32.01 ± 0.006	3.72 ± 0.020	1.19 ± 0.021	3.17 ± 0.003	3.64 ± 0.004
HXE (α=0.1)	28.41 ± 0.003	2.42 ± 0.005	0.69 ± 0.007	1.24 ± 0.005	2.24 ± 0.005
HXE (α=0.6)	30.46 ± 0.003	2.28 ± 0.009	0.69 ± 0.009	1.22 ± 0.007	2.22 ± 0.004
Soft-labels ($\beta = 30$)	27.17 ± 0.004	2.36 ± 0.001	0.64 ± 0.008	1.20 ± 0.005	2.22 ± 0.003
Soft-labels ($\beta = 4$)	32.73 ± 0.007	2.21 ± 0.023	0.72 ± 0.017	1.23 ± 0.011	2.23 ± 0.006
Chang et al. [7]	21.92 ± 0.001	2.27 ± 0.009	0.50 ± 0.003	1.10 ± 0.002	2.18 ± 0.002
HAF	22.31 ± 0.001	2.23 ± 0.018	0.50 ± 0.003	1.10 ± 0.003	2.17 ± 0.003

Table 2: Results comparing top-1 error(%) and hierarchical metrics on the test set of *CIFAR-100*. Results in the *Top* block are reported without using CRM [14] technique and *Bottom* block are reported using CRM. Rows highlighted with lightgreen are competitive methods in top-1 error (%). Of these competitive methods, we highlight the best performing entries for each metric with green.

Method	Top-1 Error(↓)	Mistakes severity(↓)	Hier dist@1(↓)	Hier dist@5(↓)	Hier dist@20(↓)
			Without CRM		
Cross-Entropy	36.44 ± 0.061	2.39 ± 0.007	0.87 ± 0.004	1.97 ± 0.002	3.25 ± 0.002
Barz & Denzler [2]	62.63 ± 0.278	1.99 ± 0.008	1.24 ± 0.005	1.49 ± 0.005	1.97 ± 0.005
YOLO-v2 [22]	44.37 ± 0.106	2.42 ± 0.003	1.08 ± 0.004	1.90 ± 0.003	2.87 ± 0.010
HXE α=0.1 [4]	41.48 ± 0.204	2.41 ± 0.009	1.00 ± 0.006	1.77 ± 0.011	2.69 ± 0.021
HXE α=0.6 [4]	45.45 ± 0.014	2.24 ± 0.006	1.02 ± 0.003	1.70 ± 0.005	2.55 ± 0.005
Soft-labels $\beta = 30$ [4]	41.67 ± 0.134	2.32 ± 0.010	0.97 ± 0.006	1.50 ± 0.006	2.23 ± 0.005
Soft-labels $\beta = 4$ [4]	74.70 ± 0.212	1.82 ± 0.005	1.36 ± 0.004	1.49 ± 0.003	1.96 ± 0.004
Chang et al. [7]	37.23 ± 0.175	2.28 ± 0.006	0.85 ± 0.004	1.75 ± 0.005	3.02 ± 0.008
HAF	36.4 ± 0.092	2.28 ± 0.012	0.83 ± 0.002	1.62 ± 0.002	2.55 ± 0.003
			With CRM		
Cross-Entropy [14]	36.51 ± 0.083	2.33 ± 0.001	0.85 ± 0.002	1.32 ± 0.001	1.86 ± 0.002
YOLO-v2	45.17 ± 0.046	2.43 ± 0.001	1.10 ± 0.001	1.50 ± 0.001	1.99 ± 0.002
HXE α=0.1	41.47 ± 0.220	2.38 ± 0.011	0.99 ± 0.008	1.41 ± 0.006	1.93 ± 0.005
HXE α=0.6	45.60 ± 0.017	2.21 ± 0.008	1.01 ± 0.003	1.40 ± 0.004	1.40 ± 0.004
Soft-labels $\beta = 30$	41.99 ± 0.126	2.31 ± 0.009	0.97 ± 0.007	1.40 ± 0.005	1.91 ± 0.005
Soft-labels $\beta = 4$	77.34 ± 0.262	2.06 ± 0.012	1.60 ± 0.007	1.72 ± 0.008	2.14 ± 0.007
Chang et al. [7]	37.31 ± 0.145	2.24 ± 0.008	0.84 ± 0.002	1.30 ± 0.002	1.84 ± 0.002
HAF	36.48 ± 0.095	2.25 ± 0.012	0.82 ± 0.003	1.29 ± 0.004	1.84 ± 0.002

Table 3: Results comparing top-1 error(%) and hierarchical metrics on the test set of *iNaturalist-19*. Results in the *Top* block are reported without using CRM [14] technique and *Bottom* block are reported using CRM. Rows highlighted with lightgreen are competitive methods in top-1 error (%). Of these competitive methods, we highlight the best performing entries for each metric with green.

Method	Top-1 error(↓)	Mistakes severity(↓)	Hier dist@1(↓)	Hier dist@5(↓)	Hier dist@20(↓)
			Without CRM		
Cross-Entropy	30.60 ± 0.030	7.05 ± 0.010	2.16 ± 0.006	5.67 ± 0.003	7.17 ± 0.003
Barz & Denzler [2]	39.73 ± 0.240	6.80 ± 0.019	2.70 ± 0.022	5.48 ± 0.271	6.21 ± 0.005
YOLO-v2 [22]	33.37 ± 0.082	7.02 ± 0.004	2.34 ± 0.016	5.85 ± 0.011	7.43 ± 0.016
DeViSE [10]	36.75 ± 0.090	6.87 ± 0.017	2.52 ± 0.009	5.57 ± 0.005	6.98 ± 0.005
HXE α=0.1 [4]	30.72 ± 0.036	7.00 ± 0.019	2.15 ± 0.005	5.62 ± 0.008	7.08 ± 0.015
HXE α=0.6 [4]	34.50 ± 0.007	6.73 ± 0.014	2.32 ± 0.003	5.48 ± 0.001	6.78 ± 0.003
Soft-labels $\beta = 30$ [4]	30.53 ± 0.194	7.05 ± 0.009	2.15 ± 0.013	5.66 ± 0.002	7.14 ± 0.008
Soft-labels $\beta = 4$ [4]	38.99 ± 0.105	6.60 ± 0.024	2.57 ± 0.004	5.13 ± 0.002	6.21 ± 0.001
Chang et al. [7]	33.46 ± 0.026	6.99 ± 0.010	2.34 ± 0.006	5.75 ± 0.005	7.34 ± 0.010
HAF	30.50 ± 0.010	7.03 ± 0.024	2.14 ± 0.008	5.62 ± 0.011	6.99 ± 0.009
			With CRM		
Cross-Entropy [14]	30.67 ± 0.020	6.99 ± 0.007	2.14 ± 0.006	4.95 ± 0.002	6.11 ± 0.001
YOLO-v2	33.98 ± 0.099	6.99 ± 0.011	2.38 ± 0.012	5.05 ± 0.001	6.17 ± 0.001
HXE α=0.1	30.80 ± 0.079	6.95 ± 0.021	2.14 ± 0.005	4.94 ± 0.003	6.11 ± 0.002
HXE α=0.6	34.68 ± 0.003	6.69 ± 0.007	2.32 ± 0.001	4.99 ± 0.005	6.13 ± 0.003
Soft-labels $\beta = 30$	30.69 ± 0.125	6.99 ± 0.007	2.15 ± 0.008	4.95 ± 0.001	6.11 ± 0.001
Soft-labels $\beta = 4$	82.72 ± 0.079	7.54 ± 0.001	6.24 ± 0.005	6.94 ± 0.005	7.25 ± 0.002
Chang et al. [7]	33.73 ± 0.033	6.93 ± 0.015	5.02 ± 0.007	2.34 ± 0.002	6.15 ± 0.001
HAF	30.63 ± 0.007	6.97 ± 0.024	2.14 ± 0.008	4.95 ± 0.004	6.11 ± 0.001

Table 4: Results comparing top-1 error(%) and hierarchical metrics on the test set of *tieredImageNet-H*. The *Top* block reports results without using CRM [14] and the *Bottom* block are reported using CRM. Rows highlighted with lightgreen are competitive methods in top-1 error (%). Of these methods, we highlight the best performing entries for each metric with green.

smaller β values, leading to predictions with low confidence, which CRM could not help rectify.

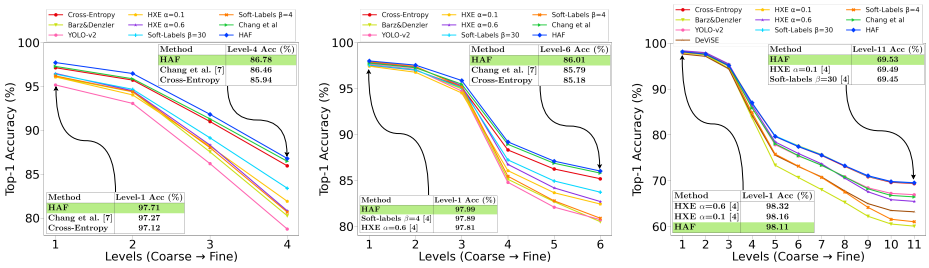
4.4 Coarse classification Accuracy

We also report comparisons over the coarse classification accuracy at all hierarchical levels. The learned feature representations guided with label hierarchies is expected to follow the structure of label hierarchies in the feature space. Such a feature space must restrict the confusions within their respective coarse classes, thereby, increasing the coarse-classification accuracy. We map the target labels and the predicted labels from the finest-level classifier to their respected coarse classes to evaluate the performance of the models on other hierarchical levels using coarse-classification accuracy. The results are reported in the Figure 3. On both CIFAR-100 and iNaturalist-19, HAF outperforms all the other baseline methods. On tieredImageNet-H, HAF has comparable performance with the Soft-labels β=30, HXE α=0.1, and HXE α=0.6.

5 Analysis

5.1 Ablation Study

In order to assess the contributions of each loss function used in our proposed approach, we present in Table 5, the results obtained with different variants of



(a) CIFAR-100 (b) iNaturalist19 (c) tieredImageNet-H

Fig. 3: Coarse-level top-1 accuracy for each dataset. Level=1 is the coarsest level.

HAF on CIFAR-100 and iNaturalist19 datasets respectively. It is evident that different variants of HAF perform slightly better than the cross-entropy baseline but HAF outperforms all the other variants. We can thus conclude that all the components of the loss function are significant and complementary for the overall performance of HAF .

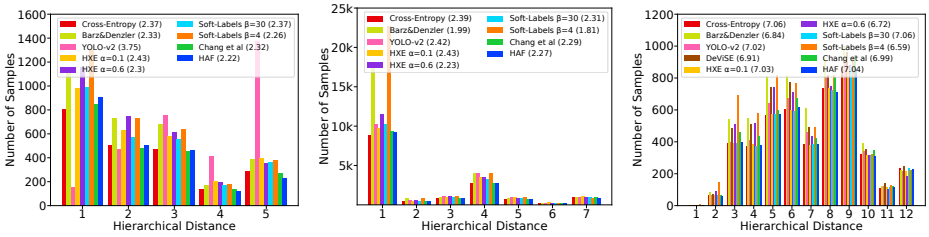
Method	Loss function $L_{CE_{fine}}$	L_{shc}	L_{gc}	L_m	Top-1 error(↓)	Mistakes severity(↓)	Hier Dist@1(↓)	Hier Dist@5(↓)	Hier Dist@20(↓)
Cross-entropy	✓	-	-	-	22.11	2.37	0.52	2.24	3.16
Variant of HAF	✓	✓	-	-	22.70	2.36	0.54	1.61	2.78
Variant of HAF	✓	✓	✓	-	22.35	2.32	0.52	1.66	2.87
Variant of HAF	✓	✓	-	✓	22.12	2.24	0.5	1.44	2.61
HAF	✓	✓	✓	✓	22.25	2.22	0.49	1.40	2.64
Cross-entropy	✓	-	-	-	36.48	2.39	0.87	1.97	3.25
Variant of HAF	✓	✓	-	-	36.23	2.34	0.85	1.73	2.81
Variant of HAF	✓	✓	✓	-	36.60	2.32	0.85	1.71	2.73
Variant of HAF	✓	✓	-	✓	36.34	2.31	0.84	1.76	2.91
HAF	✓	✓	✓	✓	36.47	2.27	0.83	1.62	2.56

Table 5: Ablative study comparing top-1 error(%) and hierarchical metrics on the test sets of *CIFAR-100* (top) and *iNaturalist-19* (bottom).

5.2 Mistakes Severity Plots

We plot histograms to compare HAF with the baselines depicting the distribution of mistakes at different hierarchical levels. We present them for each dataset in Fig. 4. Mistakes at hierarchical distance 1 refers to the mistakes with LCA=1. On CIFAR-100, HAF has the lowest mistake severity compared to all the methods and has number of mistakes comparable to cross-entropy at all levels except for level-1, where Chang et al. [7] generates fewer mistakes. However, HAF has lesser number of high severity mistakes compared to Chang et al.[7] which is a more desirable solution. On iNaturalist-19 dataset, soft-labels $\beta = 4$, Barz & Denzler, and HXE $\alpha = 0.6$ has lower mistake severity as compared to HAF , but

HAF makes lesser or nearly equal number of mistakes compared to these methods at all the hierarchical levels. On tieredImageNet-H, Barz & Denzler, DeViSE, HXE $\alpha = 0.6$, soft-labels $\beta = 4$ has lower mistakes severity than HAF but much larger number of mistakes at every level. The metric 'mistakes severity' alone does not give a complete picture of a method's ability to improve the mistakes.



(a) CIFAR-100 (b) iNaturalist19 (c) tieredImageNet-H

Fig. 4: Mistakes severity plot showing distributions of mistakes at each level for each dataset. Numbers in the bracket denote the mistake severity of the method.

5.3 Discussion: Hierarchical Metrics

We discuss the inadequacy of the hierarchical metrics that have been proposed thus far. Figure 5 plots the histogram of the smallest possible LCA for all classes of the tieredImageNet-H dataset. Most classes have a minimum LCA greater than one, which indicates a skewed hierarchy tree, in turn explaining the high values of the hierarchical metrics in Table 4 across all methods. When these metrics are averaged over samples, the resulting change turns out to be very small, as observed by the reported standard deviations in Tables 2 - 4. This problem of large values persists in all LCA-based metrics, and is dependent on the label hierarchy tree.

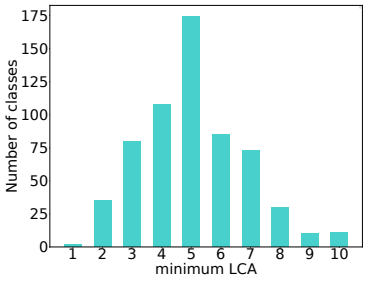


Fig. 5: Number of classes with minimum LCA for tieredImageNet-H dataset.

As is depicted above in Figure 4, mistakes severity favours the model with the reduction of *average* LCA over the mistakes, implying that this metric may prefer a model with a large number of low-severity mistakes. Karthik et. al [14] highlights the problems with mistakes severity. They overcome this drawback by using average hierarchical distance@1. However, we also note the problem with average hierarchical distance@1 metric. It is an average LCA distance of *all* the samples from ground-truth to the top-1 predictions. This average includes as

many zeros as the number of correct predictions (since the LCA distance for a correct prediction is 0). Therefore, it favours models that make fewer overall mistakes and thus fails to adequately capture the notion of a mistake's severity.

An ideal method is the one that improves the mistakes severity metric while maintaining (or improving) the top-1 error, i.e., the sum of LCA of mistakes should reduce while maintaining (or improving) the total number of errors. On the CIFAR-100 dataset, we note that with a minimal drop in the top-1 accuracy, there is nearly 5% improvement in reducing the sum of LCA of mistakes using HAF + CRM as compared to 2.17% using cross-entropy + CRM. Similarly, on iNaturalist-19, HAF + CRM minimizes the sum of LCA of mistakes by 5.68% compared to 2% on cross-entropy + CRM. We present a more detailed analysis of these metrics in the supplementary material and defer the design for a more appropriate metric to measure mistake severity for future work.

6 Conclusion

In this paper, we introduced a novel approach to learn a hierarchy-aware feature space, which can preserve or improve the top-1 error and yet reduce the severity of mistakes. Our approach uses auxiliary classifiers at each level of the hierarchy that are trained by minimizing a Jensen-Shannon Divergence with target soft labels derived from finer-grained predictions of the samples. This training strategy regularizes the fine-grained classifier to make more consistent predictions with the coarser level classifiers, leading to a reduction in severity of mistakes. We further impose geometric consistency constraints between coarse and fine classifiers that leads to better alignment of the feature space distributions of the sub-classes with that of their super-classes. Without any additional hyperparameters, we simply trained our models with these loss functions and showed a reduction in mistake severity without trading off the top-1 error. We reported results from extensive experiments over three large datasets with varying levels of hierarchy and showed the strengths of our proposed method. We also presented an analysis of the commonly used hierarchical metrics and highlighted their limitations. We note that there exist recent works that leverage non-Euclidean spaces to learn appropriate embeddings for hierarchical data. However, much of the recent work on evaluating mistake severity is restricted to Euclidean feature spaces, and we present our analysis in the same space. Nonetheless, we conjecture that the nature of our contributions in this paper, i.e., losses that impose probabilistic and geometric constraints, would also extend to non-Euclidean spaces like hyperbolic feature spaces and would serve as a promising direction for future work.

Acknowledgement

Ashima Garg was supported by SERB, Govt. of India, under grant no. CRG/2020/006049. Depanshu Sani was supported by Google's AI for Social Good "Impact Scholars" program, 2021. Saket Anand gratefully acknowledges for the partial support from the Infosys Center for Artificial Intelligence at IIIT-Delhi.

References

1. Abe, N., Zadrozny, B., Langford, J.: An iterative method for multi-class cost-sensitive learning. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 3–11 (2004)
2. Barz, B., Denzler, J.: Hierarchy-based image embeddings for semantic image retrieval. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 638–647. IEEE (2019)
3. Bengio, S., Weston, J., Grangier, D.: Label embedding trees for large multi-class tasks. In: Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., Culotta, A. (eds.) Advances in Neural Information Processing Systems. vol. 23. Curran Associates, Inc. (2010), https://proceedings.neurips.cc/paper/2010/file/06138bc5af6023646ede0e1f7c1eac75-Paper.pdf
4. Bertinetto, L., Mueller, R., Tertikas, K., Samangooei, S., Lord, N.A.: Making better mistakes: Leveraging class hierarchies with deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
5. Bilal, A., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? IEEE transactions on visualization and computer graphics **24**(1), 152–162 (2017)
6. Brust, C.A., Denzler, J.: Integrating domain knowledge: using hierarchies to improve deep classifiers. In: Asian Conference on Pattern Recognition. pp. 3–16. Springer (2019)
7. Chang, D., Pang, K., Zheng, Y., Ma, Z., Song, Y.Z., Guo, J.: Your "flamingo" is my "bird": Fine-grained, or not. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11476–11485 (2021)
8. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What does classifying more than 10,000 image categories tell us? In: European conference on computer vision. pp. 71–84. Springer (2010)
9. Deng, J., Krause, J., Berg, A.C., Fei-Fei, L.: Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 3450–3457. IEEE (2012)
10. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M.A., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 26. Curran Associates, Inc. (2013), https://proceedings.neurips.cc/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf
11. Fuglede, B., Topsoe, F.: Jensen-shannon divergence and hilbert space embedding. In: International Symposium onInformation Theory, 2004. ISIT 2004. Proceedings. p. 31. IEEE (2004)
12. Garg, A., Bagga, S., Singh, Y., Anand, S.: Hiermatch: Leveraging label hierarchies for improving semi-supervised learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1015–1024 (2022)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
14. Karthik, S., Prabhu, A., Dokania, P.K., Gandhi, V.: No cost likelihood manipulation at test time for making better mistakes in deep networks. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=193sEnKY1ij

15. Krizhevsky, A.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
16. Landrieu, L., Garnot, V.S.F.: Leveraging class hierarchies with metric-guided prototype learning. In: British Machine Vision Conference (BMVC) (2021)
17. Liu, S., Chen, J., Pan, L., Ngo, C.W., Chua, T.S., Jiang, Y.G.: Hyperbolic visual embedding learning for zero-shot recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9273–9281 (2020)
18. Mao, Y., Tian, J., Han, J., Ren, X.: Hierarchical text classification with reinforced label assignment. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 445–455 (2019)
19. Miller, G.A.: WordNet: An electronic lexical database. MIT press (1998)
20. Nickel, M., Kiela, D.: Poincaré embeddings for learning hierarchical representations. Advances in neural information processing systems **30** (2017)
21. Reddi, S.J., Kale, S., Kumar, S.: On the convergence of adam and beyond. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=ryQu7f-RZ
22. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)
23. Ren, M., Triantafillou, E., Ravi, S., Snell, J., Swersky, K., Tenenbaum, J.B., Larochelle, H., Zemel, R.S.: Meta-learning for semi-supervised few-shot classification. In: Proceedings of 6th International Conference on Learning Representations ICLR (2018)
24. Su, J., Maji, S.: Semi-supervised learning with taxonomic labels. In: British Machine Vision Conference (BMVC) (2021)
25. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
26. Wang, Y., Wang, Z., Hu, Q., Zhou, Y., Su, H.: Hierarchical semantic risk minimization for large-scale classification. IEEE Transactions on Cybernetics (2021)
27. Wehrmann, J., Cerri, R., Barros, R.: Hierarchical multi-label classification networks. In: International Conference on Machine Learning. pp. 5075–5084. PMLR (2018)
28. Wu, H., Merler, M., Uceda-Sosa, R., Smith, J.R.: Learning to make better mistakes: Semantics-aware visual food recognition. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 172–176 (2016)
29. Yang, Z., Bastan, M., Zhu, X., Gray, D., Samaras, D.: Hierarchical proxy-based loss for deep metric learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1859–1868 (2022)
30. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
31. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: Richard C. Wilson, E.R.H., Smith, W.A.P. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 87.1–87.12. BMVA Press (September 2016). https://doi.org/10.5244/C.30.87, https://dx.doi.org/10.5244/C.30.87