Hierarchical Class-Based Curriculum Loss

Palash Goyal[1] , **Divya Choudhary**[1] and **Shalini Ghosh**[2*]
[1]Samsung Research America
[2]Amazon Alexa AI
{palash.goyal, d.choudhary}@samsung.com, ghoshsha@amazon.com

Abstract

Classification algorithms in machine learning often assume a flat label space. However, most real world data have dependencies between the labels, which can often be captured by using a hierarchy. Utilizing this relation can help develop a model capable of satisfying the dependencies and improving model accuracy and interpretability. Further, as different levels in the hierarchy correspond to different granularities, penalizing each label equally can be detrimental to model learning. In this paper, we propose a loss function, hierarchical curriculum loss, with two properties: (i) satisfy hierarchical constraints present in the label space, and (ii) provide non-uniform weights to labels based on their levels in the hierarchy, learned implicitly by the training paradigm. We theoretically show that the proposed hierarchical class-based curriculum loss is a tight bound of 0-1 loss among all losses satisfying the hierarchical constraints. We test our loss function on real world image data sets, and show that it significantly outperforms state-of-the-art baselines.

1 Introduction

Machine learning (ML) models are trained on class labels that often have an underlying taxonomy or hierarchy defined over the label space. However, general ML models do not utilize the taxonomy relations between the labels and can thus make more egregious errors. For example, if an image contains "bulldog", the model would penalize a prediction of "dog" and "building" equally unlike a human evaluator who would consider "dog" to be a more accurate prediction. Although such nuances are not captured by the standard evaluation metrics, they are crucial for ensuring quality in real deployments of ML models.

Hierarchical multi-label classification (HMC) methods, which utilize the hierarchy of class labels, aim to tackle the above issue. Traditional methods in this domain broadly use one of three approaches: (i) architectural modifications to the original model to learn either levels or individual classes separately, (ii) converting the discrete label space to a continuous one and embedding the labels using relations between them, and (iii) modifying the loss function by adding more weights to specific classes in the hierarchy. However, the methods in this domain are mostly empirical and the choice of modifications is often experimental. To overcome this issue, we aim to incorporate the class dependencies in the loss function in a systematic fashion. To this end, we propose a formulation to incorporate hierarchical constraints in a base loss function and show that our proposed loss is a tight bound to the base loss.

Further, we note that typically humans do not learn all the categories of objects at the same time, but rather learn them gradually starting with simple high-level categories. A similar setting was explored by Bengio *et al.* [Bengio *et al.*, 2009], introducing the concept of curriculum learning feeding the model easier examples to mimic the way of human learning. They show that learning simple examples first makes the model learn a smoother function. Lyu *et al.* [Lyu and Tsang, 2019] extended this to define an example-based curriculum loss with theoretical bounds to 0-1 loss. We extend our hierarchically constrained loss function to incorporate a class-based curriculum learning paradigm, implicitly providing higher weights to simpler classes. With the hierarchical constraints, the model ensures that the classes higher in the hierarchy are selected to provide training examples until the model learns to identify them correctly, before moving on to classes deeper in the hierarchy (making the learning problem more difficult).

We theoretically show that our proposed loss function, hierarchical class-based curriculum loss, is a tight bound on 0-1 loss — we show that any other loss function that satisfies hierarchical constraints on a given base loss gives a higher loss compared to our loss. We evaluate this result empirically on four image data sets, showing that our loss function provides a significant improvement on the hierarchical distance metric compared to the baselines. We also show that, unlike many other hierarchical multi-label classification methods, our method does not have a degraded performance on non-hierarchical metrics and in most cases has a significant improvement over the baselines.

2 Related Work

Research in hierarchical classification falls into three categories: (i) embedding label relations in continuous space, (ii) structural modifications of base architecture, and (iii) adding hierarchical regularizers and loss function modifications.

*The author only contributed to part of this work while working at Samsung Research America.

Label-embedding methods [Kumar *et al.*, 2018; Frome *et al.*, 2013; Chen *et al.*, 2019; He and Chua, 2017; Huang and Lin, 2017; Bertinetto *et al.*, 2019] map class labels to continuous vectors capable of reconstructing the relation between labels. They then learn a model to predict the embedding instead of original labels. The disadvantage of these methods is typically the difficulty of mapping back the prediction to the discrete space and the noise introduced in this conversion.

Models that perform structural modifications use earlier layers in the network to predict higher level categories and later layers to predict lower level categories [Wehrmann *et al.*, 2018; Masera and Blanzieri, 2018; Cerri *et al.*, 2014; Bilal *et al.*, 2017]. Structural methods are often domain-dependent, needing time and effort to analyze the data and come up with specific modifications. In comparison, our loss based HCL method is easier to implement and domain-agnostic.

Finally, models that modify the loss function to incorporate hierarchy assign a higher penalty to the prediction of labels which are more distant from the ground truth. AGGKNN-L and ADDKNN-L-CSL [Verma *et al.*, 2012] use a lowest common ancestor (LCA) based penalty between the classes. Similarly, Deng *et. al.* [Deng *et al.*, 2010] introduce hierarchical cost giving penalty based on the height of LCA. CNN-HL-LI [Wu *et al.*, 2016] use a weighting parameter to control the contribution of fine-grained classes which is empirically learned. HXE [Bertinetto *et al.*, 2019] use a probabilistic approach to assign penalties for a given class given the parent class and provide an information theoretic interpretation for it, while Papai et al. [Papai *et al.*, 2012] use an approach where a subjective feedback provided by a user is translated to probabilistic constraints.

The domain of curriculum learning was introduced by Bengio *et. al.* [Bengio *et al.*, 2009] based on the observation that humans learn much faster when presented information in a meaningful order as opposed to random, which is typically used for training machine learning models. Several follow up works have shown this type of learning to be successful [Smith, 2017; Matiisen *et al.*, 2019; Lyu and Tsang, 2019]. Our approach builds on this line of work to develop a curriculum learning-based loss.

3 Preliminaries

We first define some notations for the rest of the paper. Let $\mathbb{T} = (\boldsymbol{x}_i, \boldsymbol{y}_i)_{i=1,\ldots,N}$ denote the training set with N examples where $\boldsymbol{x}_i \in \mathbb{R}^D$ is the input feature vector and $\boldsymbol{y}_i \in \{0,1\}^C$ is the ground truth assignment of the input to a category space. Here, C and D are the number of categories and dimensionality of input feature vector respectively. Note that we consider the problem of multi-label classification where an input can belong to multiple classes simultaneously. We denote the loss function as $l : \mathbb{R}^N \times \mathbb{R}^C \to \mathbb{R}$ where $l(\boldsymbol{y}_i, \hat{\mathbf{y}}_\mathbf{i})$ denotes the loss when the prediction is $\hat{\mathbf{y}}_\mathbf{i}$. For ease of notation, we will refer to it as $l(\hat{\mathbf{y}}_\mathbf{i})$. In our problem, we assume that the set of categories $\mathbb{C}$ can be arranged in a hierarchy $\mathcal{H} = (\mathbb{C}, \mathbb{P})$ where $\mathbb{P}$ is the set of directed parent-child relations between the categories, $\mathbb{P}(c) = p$ means parent of class c is p.

4 Hierarchical Class-Based Curriculum Loss

For the task of multi-class classification, given a multi-class loss function, our goal is to incorporate the hierarchical constraints present in the label space into the given loss function. In this section, we first define a hierarchical constraint which we require to be satisfied for a hierarchical label space. We then introduce our formulation of a hierarchically constrained loss and show that the proposed loss function indeed satisfies the constraint. We prove a bound on the proposed loss, extend the loss function using a curriculum learning paradigm to get a tight bound to the 0-1 loss using this, and present an algorithm to train the model using the proposed loss function. Note that the proposed loss function HCL corresponds to l_{hc}. HCL-Hier (l_h) is any generic loss that satisfies our proposed hierarchical constraint, the first component of HCL. HCL-CL (l_{hc} without l_h) is any generic loss that just uses class-based curriculum learning, the second component of HCL.

4.1 Incorporating Hierarchical Constraints

Consider the learning framework with training set $\mathbb{T} = (x_i, y_i)_{i=1,\ldots,N}$. A general multi-class multi-label loss lower bounded by 0-1 loss can be defined as follows:

$$l(\hat{y}) = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} l(\hat{y}_{i,j}). \tag{1}$$

Note that this loss function does not impose any hierarchical constraints between the categories. Consider two categories in the category set $\mathbb{C}$: *Animal* and *Dog*. Without constraints, a model trained using the above loss can have $l_{Animal} > l_{Dog}$. This implies that the model is more accurate in its prediction of *Dog* compared to *Animal*. However, this is counter-intuitive as *Dog* is an *Animal* and its confidence should not be greater than *Animal*'s. To tackle this, we define the following hierarchical constraint Λ on a generic loss function l:

$$\begin{aligned} &\Lambda : \forall (x_i, y_i) \in \mathbb{T}, \forall c_1, c_2 \in \mathbb{C}, \\ &\mathbb{P}(c_1) = c_2 \implies l(\hat{y}_{i,c_1}) \geq l(\hat{y}_{i,c_2}) \end{aligned} \tag{2}$$

The constraint implies that the loss of a child node in the hierarchy is constrained to be more than the loss of the corresponding parent node. Satisfaction of this constraint encodes that the model suffers a lower loss when it predicts a finer category with less accuracy than its parent, which is intuitive. Note that this constraint can be modified to constrain level-wise relations instead of parent-child relations. The theorems in this paper can be proven for level-wise relations as well.

We now propose the loss function l_h and show that it satisfies the hierarchical constraint Λ by definition:

$$l_h(\hat{y}) = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} \max(l(\hat{y}_{i,j}), l(\hat{y}_{i,\mathbb{P}(j)})) \tag{3}$$

We now show the following result for l_h.

Theorem 1. ***(Hierarchically Constrained Loss)*** *Given a loss function l, the loss function l_h defined in Equation 3 satisfies the hierarchical constraint Λ defined in Equation 2.*

Proof. Let us assume that l_h doesn't follow hierarchical constraint Λ i.e. $\exists (x_i, y_i) \in \mathbb{T}, \exists c_1, c_2 \in \mathbb{C}$ s.t.

$$\mathbb{P}(c_1) = c_2 \text{ and } l_h(\hat{y}_{i,c_1}) < l_h(\hat{y}_{i,c_2})$$

However, we have

$$l_h(\hat{y}_{i,c_1}) = \max(l(\hat{y}_{i,c_1}), l(\hat{y}_{i,\mathbb{P}(c_1)})) \geq l_h(\hat{y}_{i,c_2})$$

which contradicts our assumption. This implies that l_h has to follow the hierarchical constraint Λ. □

Now, we show that the hierarchically constrained loss function is tightly bounded to the base function.

Theorem 2. ***(Bound on Constrained Loss)*** *For a loss function $l(\hat{y})$ lower-bounded by 0-1 loss, the loss function l_h defined in Equation 3 is an element-wise tight bound on $l(\hat{y})$ with constraint Λ. Let $\preceq$ denote elementwise inequality i.e. $f \preceq q$ means $\forall x \in domain(f), f(x) \leq q(x)$. We then have:*

$$l \preceq l_h \preceq g \; \forall g \in \mathcal{L} \text{ satisfying constraint } \Lambda, \text{ s.t. } l \preceq g. \quad (4)$$

Proof. Let us assume that $\exists g \in \mathcal{L}$ s.t. $l_h \not\preceq g$ i.e.

$$\exists (x_i, y_i) \in \mathbb{T}, \exists c_1 \in \mathbb{C} \text{ s.t. } g(\hat{y}_{i,c_1}) < l_h(\hat{y}_{i,c_1}) \quad (5)$$

Using the definition of l_h from Eq. 3, we have:

$$g(\hat{y}_{i,c_1}) < max(l(\hat{y}_{i,c_1}), l(\hat{y}_{i,\mathbb{P}(c_1)})) \quad (6)$$

This leads to two cases:

Case 1. If $l(\hat{y}_{i,c_1}) > l(\hat{y}_{i,\mathbb{P}(c_1)})$, then we get $g(\hat{y}_{i,c_1}) < l(\hat{y}_{i,c_1})$. This contradicts the assumption that g is lower-bounded by l i.e. $l \preceq g$.

Case 2. If $l(\hat{y}_{i,\mathbb{P}(c_1)}) > l(\hat{y}_{i,c_1})$, then we get $g(\hat{y}_{i,c_1}) < l(\hat{y}_{i,\mathbb{P}(c_1)})$. As $l \preceq g$, we have $g(\hat{y}_{i,\mathbb{P}(c_1)}) >= l(\hat{y}_{i,\mathbb{P}(c_1)})$. Combined, we get $g(\hat{y}_{i,\mathbb{P}(c_1)}) > g(\hat{y}_{i,c_1})$. However, this violates the Λ constraint.

Thus, we get $l \preceq l_h \preceq g$. □

The above result is for a generic loss — it can be used to get the following property for 0-1 loss.

Corollary 1. ***(Bound on Constrained 0-1 Loss)*** *For a 0-1 loss function $e(\hat{y})$, the loss function e_h is defined as*

$$e_h(\hat{y}) = \sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} \max(e(\hat{y}_{i,j}), e(\hat{y}_{i,\mathbb{P}(j)})) \quad (7)$$

is an element-wise tight bound on $e(\hat{y})$ with constraint Λ, i.e., the following relation holds:

$$e \preceq e_h \preceq g \; \forall g \in \mathcal{L} \text{ satisfying constraint } \Lambda, \text{ s.t. } e \preceq g. \quad (8)$$

We have shown above that the hierarchy constrained loss function l_h provides an element-wise tight bound on the base loss l. We now extend this loss function to use a curriculum learning paradigm and show that the loss is a tighter bound to 0-1 loss compared to any other hierarchy preserving loss.

4.2 Hierarchical Curriculum Loss

As shown by Hu et al. [Hu *et al.*, 2016], 0-1 loss ensures that the empirical risk has a monotonic relation with adversarial empirical risk. However, it is non-differentiable and difficult to optimize. Following the groundwork by Lyu et al. [Lyu and Tsang, 2019] who propose example based curriculum loss, we present a class-based curriculum loss for any given hierarchically constrained loss function l and a class selection parameter $\boldsymbol{s}$ in the following theorem. The theorem also proves that the function defined is tighter bound to 0-1 loss compared to any loss function which satisfies the hierarchical constraint and is lower bounded by l. Note that a general loss function l is element-wise lower bounded by 0-1 loss e, i.e., $e \preceq l$.

Theorem 3. ***(Hierarchical Class-Based Curriculum Loss)*** *For a general hierarchy constrained loss function $l_h(\hat{y})$, we define the loss function $l_{hc}(\hat{y})$ as follows:*

$$\min_{s \in \{0,1\}^C} \max \left(\sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} s_j l_h(\hat{y}_{i,j}), C - \sum_{j \in \mathbb{C}} s_j + e_h(\hat{y}) \right) \quad (9)$$

Then, $e(\hat{y}) \leq l_{hc}(\hat{y}) \leq g(\hat{y}) \; \forall g \in \mathcal{L}$ satisfying hierarchical constraint Λ, s.t. $l \preceq g$ i.e. the following holds

$$|l_{hc}(\hat{y}) - e(\hat{y})| \leq |g(\hat{y}) - e(\hat{y})| \forall g \in \mathcal{L} \text{ such that} \quad (10)$$

$$\begin{aligned} &\forall (x_i, y_i) \in \mathbb{T}, \forall c_1, c_2 \in \mathbb{C}, \\ &\mathbb{P}(c_1) = c_2 \implies g(\hat{y}_{i,c_1}) \geq g(\hat{y}_{i,c_2}) \end{aligned} \quad (11)$$

$$l(\hat{y}_{i,j}) \leq g(\hat{y}_{i,j}) \forall (x_i, y_i) \in \mathbb{T}, \forall j \in \mathbb{C} \quad (12)$$

Proof. Consider $l_{hc}(\hat{y})$ which is

$$\begin{aligned} &\min_{s \in \{0,1\}^C} \max(\sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} s_j l_h(\hat{y}_{i,j}), C - \sum_{j \in \mathbb{C}} s_j + e_h(\hat{y})) \\ &\leq \max(l_h(\hat{y}), e_h(\hat{y})) \\ &= l_h(\hat{y}) \leq g(\hat{y}) \text{ (from Theorem 2).} \end{aligned} \quad (13)$$

For the lower bound we have,

$$\begin{aligned} &\min_{s \in \{0,1\}^C} \max(\sum_{i \in \mathbb{N}} \sum_{j \in \mathbb{C}} s_j l_h(\hat{y}_{i,j}), C - \sum_{j \in \mathbb{C}} s_j + e_h(\hat{y})) \\ &\geq \min_{s \in \{0,1\}^C} (C - \sum_{j \in \mathbb{C}} s_j + e_h(\hat{y})) = e_h(\hat{y}) \\ &\geq e(\hat{y}) \text{(from Corollary 1).} \end{aligned} \quad (14)$$

This gives $e(\hat{y}) \leq l_{hc}(\hat{y}) \leq g(\hat{y})$. Subtracting $e(\hat{y})$ from both sides, we get the theorem. □

Note that $\boldsymbol{s}$ governs classes to be selected for training, i.e., $\boldsymbol{s}(i) = 1$ implies that class i is selected by the loss function. This is primarily used to select easy classes first followed by harder classes. Intuitively, the max in the loss function sub-selects classes based on their loss values.

4.3 Algorithm

In the above theorem, we prove that the proposed hierarchical class-based curriculum loss provides a tighter bound to 0-1 loss compared to the hierarchically constrained loss function. Given the above, we now need to find the optimal class selection parameters s_i for each class. We show that Algorithm 1 provides the optimal selection parameters.

Theorem 4. ***(Class Selection)*** *Given a base loss function l, a hierarchically constrained loss function l_h, a solution s for Equation 9 is provided by Algorithm 1.*

The proof for above is provided in the Appendix. The algorithm first creates a hierarchically constrained loss function l_h given the loss function l. It then sorts the loss values of classes in increasing order of magnitude. Finally, it selects the first K classes from this list such that the cumulative sum is greater than $thresh + 1 - K$, where $thresh$ is an hyper-parameter. The first $K - 1$ classes go in the selection pool.

Note that the time complexity of our approach is O(NClog(C)) and that of a typical base loss is O(NC). The additional log(C) is computationally inexpensive as class sizes are typically very small.

Algorithm 1: Class Selection for Hierarchical Class-Based Curriculum Learning. Number of samples N, Number of classes C, Base Loss l, Threshold $thresh$- determines the selection of classes to contribute to the loss function in an iteration

Function *selectClasses (l, thresh)*
- **for** $j = 1 \dots C$ **do**
  - $l_h(\hat{y}_{.,j}) \leftarrow 0$;
  - **for** $i = 1 \dots N$ **do**
    - $l_h(\hat{y}_{i,j}) \leftarrow \max(l(\hat{y}_{i,j}), l(\hat{y}_{i,\mathbb{P}(j)}))$;
    - $l_h(\hat{y}_{.,j})$ += $l_h(\hat{y}_{i,j})$;
- Sort class indices in non-decreasing order of l_h;
- Get minimum K s.t. $\sum_{c=1}^{K} l_h(\hat{y}_{.,c}) > thresh + 1 - K$;
- **for** $i = 1 \dots C$ **do**
  - **if** $i < K$ **then**
    - $s_i \leftarrow 1$;
  - **else**
    - $s_i \leftarrow 0$;
- return s

5 Experiments

In this section, we present experiments to showcase the performance of hierarchical class-based loss. We show the results on four data sets with the state-of-the-art hierarchical loss functions. Further, we perform statistical tests to show the significance of our results. We also perform an ablation study on each of the component of hierarchical class-based curriculum loss, including the hierarchically constrained loss(HCL-Hier) and class-based curriculum loss(HCL-CL), and show how they interplay to provide the proposed final loss function(HCL). We performed our experiments on 2 Nvidia GeForce RTX 2080 Ti with 12 GB memory with 3.30 GHz CPU clock speed.

5.1 Experimental Setup

We evaluate our loss function on four real world image data sets – (i) IMCLEF [Dimitrovski *et al.*, 2011], (ii) Wipo [Rousu *et al.*, 2006], (iii) Reuters [Lewis *et al.*, 2004], and (iv) iNaturalist [Van Horn *et al.*, 2018]. A summary of these datasets is provided in Tables 1 & 2.

Experimental Details. For evaluation on datasets with pre-extracted features, we use a multi-layer perceptron with the extracted features as input and the categories as output. We select the hyperparameters of the neural network using evaluation on a validation set with binary cross entropy loss. Based on this, we get a structure with 800 hidden neurons and a dropout of 0.25. Note that we fix this network for all the baseline loss functions and our loss function to ensure fair comparison of results. For evaluation on iNaturalist, we used a ResNet-18 architecture (pre-trained on ImageNet). We use Adam optimizer and a learning rate of 10^{-5}.

Baselines. We compare the hierarchical class-based curriculum loss with the following state-of-the-art losses – (i) binary cross entropy loss [Goodfellow *et al.*, 2016], (ii) focal loss [Lin *et al.*, 2018] and (iii) hierarchical cross entropy loss [Bertinetto *et al.*, 2019]. Further, we also compare it with a label-embedding approach called SoftLabels [Bertinetto *et al.*, 2019], which modifies the ground truth labels in accordance with the hierarchy. For all the baselines, we use the range of hyperparameters used in their respective works.

5.2 Metrics

We use the following metrics to evaluate each of the losses for the classification task – (i) Hit@1, (ii) MRR (Mean Reciprocal Rank) [Radev *et al.*, 2002], (iii) HierDist [Deng *et al.*, 2010], (iv) HierCons. The first two metrics capture the accuracy of ranking of the model predictions while the other two metrics capture preservation of hierarchical information. Hierarchy capturing methods often show lower performance compared to non-hierarchical methods on first two metric as the losses get more constrained. However, these hierarchical methods often show improvements on a metric which captures how close to the ground truth class the prediction is in the given hierarchy.

HierDist captures this and is defined as the minimum height of the lowest common ancestor (LCA) between the ground truth labels and the top prediction from the model. Mathematically, for a data point $(x_i, y_i) \in \mathcal{T}$, it is defined as

$$HierDist = \min_{c_1 \in \{j : y_{i,j} == 1\}} LCA_{\mathbb{H}}\big(c_1, argmax_j(\hat{y}_{i,j})\big),$$

where $\mathbb{H}$ denotes the hierarchy of the labels. The minimum height of LCA is taken as 0 for a correct prediction. As pointed out by Deng [Deng *et al.*, 2010], the metric is effectively on a log scale. It is measured by the height in the hierarchy of the lowest common ancestor, and moving up a level can be more than double the number of descendants depending on the fan out of the parent class (often greater than 3-4). We show that our loss function is superior to the baseline losses for this metric. In addition, our model's performance also doesn't deteriorate on non-hierarchical metrics.

We further propose another metric Hierarchical Consistency (HierCons) to measure the consistency of predictions w.r.t. the hierarchy. We define a pair of predictions to be consistent if they satisfy constraint Λ (Eq. 2). HierCons@k considers the top k predictions and takes the ratio of total consistent pairs and maximum possible pairs which is ${}^{k}C_2$.

5.3 Overall Results

We now compare our proposed loss (HCL) with the state-of-the-art loss functions capturing hierarchy as well as a label embedding method (SoftLabels). From Table 1 & 2, we observe that our loss significantly outperforms the base loss functions. Consistent with earlier results [Bertinetto *et al.*, 2019], we see that previously proposed hierarchically constrained loss functions especially SoftLabels & HXE improve on hierarchical metrics but the performance deteriorates for non-hierarchical metrics. On the other hand, HCL is able to get significant boost in Hit@1, MRR and HierDist while also improving or maintaining similar HierCons@3 as compared to all state-of-the-art loss functions including the base cross-entropy loss. HierCons@3 is also a dataset dependent metric, iNaturalist data sets have higher inherent consistency than say IMCLEF. We observe that overall Hier-CE is the best performing baseline but our model outperforms it on every

Methods	Wipo (T_r/T_t:1352/358)				Reuters (T_r/T_t:3000/3000)				IMCLEF (T_r/T_t:10000/1006)				
	Hit@1	MRR	HierDist	HierCons@3	Hit@1	MRR	HierDist	HierCons@3	Hit@1	MRR	HierDist	HierCons@3	Time
CE	82.91	87.71	0.510	0.949	97.90	98.76	0.084	0.937	90.35	92.97	0.285	0.765	6.32 sec
FL	79.55	84.78	0.602	0.940	97.63	98.59	0.093	0.912	90.14	93.76	0.263	0.764	9.11 sec
HXE	82.91	87.39	0.513	0.951	97.80	98.65	0.085	0.889	90.45	93.27	0.279	0.766	93.05 sec
SL	79.83	87.18	0.605	**0.993**	97.93	98.90	0.083	**0.997**	89.95	**93.85**	0.300	**0.939**	6.29 sec
HCL	**85.43**	**89.15**	**0.434**	0.944	**98.17**	**98.90**	**0.073**	0.981	**91.24**	93.66	**0.258**	0.834	10.31 sec

Table 1: Hierarchical Image Classification Results on Wipo, Reuters and IMCLEF data sets. T_r, T_t represent number of training and test examples respectively. We use pre-extracted features with a multi-layer perceptron as our base model. We show time taken per epoch for training for the larger data set IMCLEF.

Methods	Reptiles (T_r/T_t:12357/2158)				Fungi (T_r/T_t:1931/355)				Birds (T_r/T_t:40619/7250)				
	Hit@1	MRR	HierDist	HierCons@3	Hit@1	MRR	HierDist	HierCons@3	Hit@1	MRR	HierDist	HierCons@3	Time
CE	95.22	96.51	0.334	0.971	81.36	87.88	1.305	1.0	95.72	96.98	0.299	0.974	5.44 sec
FL	95.36	96.79	0.322	0.975	81.92	87.82	1.266	1.0	96.15	97.33	0.266	0.940	6.29 sec
HXE	95.09	**97.30**	0.344	**0.995**	81.64	87.92	1.285	1.0	93.03	96.30	0.488	0.993	6.67 sec
SL	95.22	95.83	0.329	0.953	76.84	83.69	1.621	1.0	96.59	96.93	0.236	0.960	5.85 sec
HCL	**95.64**	96.77	**0.305**	0.993	**83.33**	**89.13**	**1.167**	1.0	**96.68**	**97.58**	**0.230**	**0.999**	5.60 sec

Table 2: Hierarchical Image Classification Results on iNaturalist data set. We use ResNet-18 as our base model.We show time taken per epoch for training for the larger iNaturalist data set Birds.

Methods	Average Ranking in Metric across Data Sets				
	Hit@1	MRR	HierDist	HierCons@3	Overall
CE	3.16	3.16	3.5	3.00	3.20
FL	3.33	3.16	3.0	3.83	3.33
HXE	3.33	3.00	3.83	2.33	3.12
SL	3.33	3.33	3.66	2.16	3.13
HCL	**1.00**	**1.66**	**1.00**	**2.00**	**1.42**

Table 3: The rankings of HCL and baselines obtained using Tables 1 and 2. For each metric, the ranking is averaged over all the data sets. Overall ranking takes average over the metrics.

Methods	Hit@1	MRR	HierDist	HierCons@3	DevHit@1
CE	90.35	92.97	0.285	0.765	0.430
HCL-Hier	90.45	93.27	0.283	0.769	0.322
HCL-CL	90.75	93.41	0.274	**0.838**	0.416
HCL	**91.24**	**93.66**	**0.258**	0.834	0.353

Table 4: Ablation studies on IMCLEF showing the effect of each component of our loss function. For HierDist & DevHit@1, lower value is better.

metric except HierCons@3 on Wipo. Label embedding of SoftLabels(SL) facilitate parent classes to contribute to loss of all children classes in the hierarchy, thus inherently preserving hierarchical consistency. HCL significantly outperforms SL in all metric except HierCons@3 with slight trade-off on MRR in IMCLEF. Note that HCL is also computationally efficient and takes the least time compared to the best performing hierarchical baselines (e.g. HXE). We ranked performance of HCL loss as compared to other state-of-the-art losses on all 4 metric across datasets. As per Table 3, HCL has the best rank in terms of improvement across metrics, slightly above 1 is contributed by slightly lesser improvement in MRR with sizeable gains in other metrics. We present statistical tests on this table in the next section.

5.4 Detailed Analysis

We perform an ablation study on each component of the proposed loss and show their interplay. We then show statistical analyses to further evaluate the performance of the final loss, HCL. Finally, we show how the hyperparameter of HCL was selected and its effect on performance.

Ablation Studies

We show the effects of the hierarchical constraints and the curriculum loss using cross entropy loss as the base function in Table 4. We show results on test set of IMCELF.

We observe that the hierarchically constrained cross entropy loss, HCL-Hier, improves on the hierarchical metrics and non-hierarchical metrics. This shows that adding the hierarchical constraint doesn't deteriorate typical metrics all the while making the predictions more consistent with the label space hierarchy as well as making predictions closer to the ground truth label in the hierarchical space. We believe that the improvement in non-hierarchical metrics may be attributed to the modified ranking due to hierarchical constraint making the top predictions more likely.

Looking into class-based curriculum loss individually, we observe that as suggested by theory, carefully selecting the classes based on their loss (giving more weight to simpler classes at higher hierarchy levels) improves the hierarchical evaluation metric, making the predictions more relevant.

Overall, combining the two aspects yields our proposed loss HCL which shows the best performance. We see that for all the metrics, HCL gives significant gains both with respect to baseline loss and individual components. This follows theory in which we have shown that combining class-based curriculum loss with hierarchical constraints gives a tighter bound to 0-1 loss with respect to the hierarchically constrained loss. Further, as this loss explicitly ensures that loss of a higher level node is lower than the a lower level, and implicitly gives more weight to higher level node as they are selected more, the combined effect makes it more consistent to the hierarchy. This can be particularly seen from the fact that

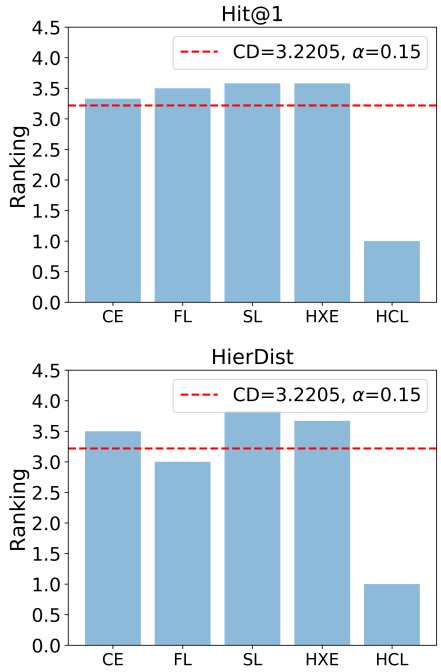

Figure 1: Statistic graph of Bonferroni-Dunn test corresponding to the results reported in Tables 1 & 2. The red dashed line denotes the sum of ranking of HCL and corresponding CD. The algorithms above the line are statistically significantly outperformed by HCL.

each individual component gave good improvements for the hierarchical metric but the combined loss gave much more significant gain. HCL loss finds the best trade-off between maintaining hierarchical metrics HierDist & HierCons while significantly improving non-hierarchical metrics Hit@1 & MRR. It is able to improve Hit@1, MRR and even HierDist much significantly with much lower deviation in Hit@1 metric compared to base cross-entropy loss when both hierarchical constraints and curriculum loss are applied together.

Quantitative Analysis

We perform statistical tests to understand the significance of our results. Based on Tables 1 & 2, the methods are ranked (summarized in Table 3). We evaluate our method using Friedman test and Bonferroni-Dunn's test.

Based on Friedman test, we get a χ^2 value of 12.1 and 12.93 for Hit@1 and HierDist respectively (with p-values 0.017 and 0.011 respectively). Thus, the results of various approaches vary widely with a confidence value of 95% enabling us to use difference statistics to compare the approaches.

We use Bonferroni-Dunn's test to identify the difference between the methods. The critical difference (CD_α) is given by the formula $q_\alpha \sqrt{\frac{g(g+1)}{6N}}$, where q_α is the critical value, g is the number of methods and N is the number of samples (i.e. data sets in our case). For $\alpha = 0.15$, we have q_α of 2.4324 which leads to CD_α of 2.2205. The comparison of methods with HCL is provided in Figure 1. We conclude that with 85% confidence, HCL outperforms all baselines on Hit@1 metric and all methods except Focal Loss on HierDist metric.

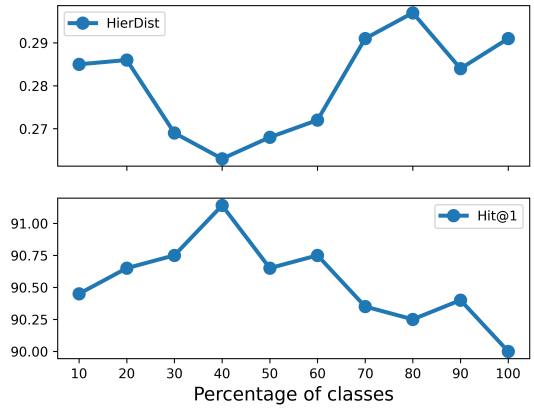

Figure 2: Evaluation of varying percentage class-selection threshold for HCL. Results are shown for HierDist (top) and Hit@1 (bottom).

Hyperparameter Variation

As mentioned in Algorithm 1, HCL uses a threshold which determines the selection of classes to contribute to the loss function at a particular iteration. We select the appropriate value of the threshold based on the percentage of classes selected with the random initialization of the model (i.e. using the 0th iteration loss values). We analyze the impact of varying the percentage of class-selection by HCL on its performance. From Figure 2 we observe that, both significant metrics, HierDist and Hit@1 change significantly in their performance based on the percentage of classes selected and hence threshold used for class-selection. HierDist is at its best of 0.258 and Hit@1 at its best of 91.24 at 40% class-selection of the total 47 classes in IMCLEF. We also note that the drop in the performance with less than optimal or more than the required class-selection percentage is smooth and in accordance with the learning-method of HCL loss as explained in section 4.2.

6 Conclusion

In this paper, we propose a hierarchical class-based curriculum loss for multi-label classification with theoretical analysis and provable bounds, making it general and dataset-agnostic. We propose a class-based curriculum loss to enhance the performance of the hierarchically constrained loss, and show significant empirical gains on multiple data sets. We observed that our models improve on both hierarchical and non-hierarchical metrics, making it widely applicable. We further perform statistical tests to illustrate the significance of results.

In the future, we would like to relax the hierarchical constraints and develop a loss for a general graph-based relation structure between the labels. Further, we would like to test the model on other real world data sets that contain relation between labels. Finally, we would also like to test the model performance when we introduce noise into the hierarchical labels of the class taxonomy.

References

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In

Proceedings of the 26th annual international conference on machine learning, pages 41–48, 2009.

[Bertinetto *et al.*, 2019] Luca Bertinetto, Romain Mueller, Konstantinos Tertikas, Sina Samangooei, and Nicholas A Lord. Making better mistakes: Leveraging class hierarchies with deep networks. *arXiv preprint arXiv:1912.09393*, 2019.

[Bilal *et al.*, 2017] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics*, 24(1):152–162, 2017.

[Cerri *et al.*, 2014] Ricardo Cerri, Rodrigo C Barros, and André CPLF De Carvalho. Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, 80(1):39–56, 2014.

[Chen *et al.*, 2019] Chen Chen, Haobo Wang, Weiwei Liu, Xingyuan Zhao, Tianlei Hu, and Gang Chen. Two-stage label embedding via neural factorization machine for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3304–3311, 2019.

[Deng *et al.*, 2010] Jia Deng, Alexander C Berg, Kai Li, and Li Fei-Fei. What does classifying more than 10,000 image categories tell us? In *European conference on computer vision*, pages 71–84. Springer, 2010.

[Dimitrovski *et al.*, 2011] Ivica Dimitrovski, Dragi Kocev, Suzana Loskovska, and Sašo Džeroski. Hierarchical annotation of medical images. *Pattern Recognition*, 44(10-11):2436–2449, 2011.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, 2013.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[He and Chua, 2017] Xiangnan He and Tat-Seng Chua. Neural factorization machines for sparse predictive analytics. In *Proceedings of the 40th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 355–364, 2017.

[Hu *et al.*, 2016] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? *arXiv preprint arXiv:1611.02041*, 2016.

[Huang and Lin, 2017] Kuan-Hao Huang and Hsuan-Tien Lin. Cost-sensitive label embedding for multi-label classification. *Machine Learning*, 106(9-10):1725–1746, 2017.

[Kumar *et al.*, 2018] Vikas Kumar, Arun K Pujari, Vineet Padmanabhan, Sandeep Kumar Sahu, and Venkateswara Rao Kagita. Multi-label classification using hierarchical embedding. *Expert Systems with Applications*, 91:263–269, 2018.

[Lewis *et al.*, 2004] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5(Apr):361–397, 2004.

[Lin *et al.*, 2018] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[Lyu and Tsang, 2019] Yueming Lyu and Ivor W Tsang. Curriculum loss: Robust learning and generalization against label corruption. *arXiv preprint arXiv:1905.10045*, 2019.

[Masera and Blanzieri, 2018] Luca Masera and Enrico Blanzieri. Awx: An integrated approach to hierarchical-multilabel classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 322–336. Springer, 2018.

[Matiisen *et al.*, 2019] Tambet Matiisen, Avital Oliver, Taco Cohen, and John Schulman. Teacher-student curriculum learning. *IEEE transactions on neural networks and learning systems*, 2019.

[Papai *et al.*, 2012] Tivadar Papai, Shalini Ghosh, and Henry A. Kautz. Combining subjective probabilities and data in training markov logic networks. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD*, 2012.

[Radev *et al.*, 2002] Dragomir R Radev, Hong Qi, Harris Wu, and Weiguo Fan. Evaluating web-based question answering systems. In *LREC*, 2002.

[Rousu *et al.*, 2006] Juho Rousu, Craig Saunders, Sandor Szedmak, and John Shawe-Taylor. Kernel-based learning of hierarchical multilabel classification models. *Journal of Machine Learning Research*, 7(Jul):1601–1626, 2006.

[Smith, 2017] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 464–472. IEEE, 2017.

[Van Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[Verma *et al.*, 2012] Nakul Verma, Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. Learning hierarchical similarity metrics. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2280–2287. IEEE, 2012.

[Wehrmann *et al.*, 2018] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In *International Conference on Machine Learning*, pages 5075–5084, 2018.

[Wu *et al.*, 2016] Hui Wu, Michele Merler, Rosario Uceda-Sosa, and John R Smith. Learning to make better mistakes: Semantics-aware visual food recognition. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 172–176, 2016.