1 Hierarchical cross-entropy (HXE) [2]

Here, we provide a derivation asserting our claims in the main text that the hierarchical cross-entropy (HXE) loss is a weighted combination of the cross-entropy loss applied at different hierarchical levels.

Following the notations as given in [2]. Define $p(C)$ is the categorical distribution over classes. The path from a leaf node C to the root R is $C^{(0)} = C, \ldots, C^{(h)} = R$, the probability of class C can be factorised as

$$p(C) = \prod_{l=0}^{h-1} p(C^{(l)}|C^{(l+1)}) \tag{1}$$

Level h=0 is the finest-level in [2]. Therefore, $p(C^{(h)}) = 1$ at root level, and hence, last term is omitted in the above expression. Further, we use L^{h}_{CE} to denote the cross-entropy at level-h. y_{true} is equal to 1 if the true class is same as that of C^{k}.

$$L^{0}_{\text{CE}}(p, C) = -y_{\text{true}} * \log p(C) \tag{2}$$

$$L^{0}_{\text{CE}}(p, C) = -y_{\text{true}} * \log \big(p(C^{0}|C^{1}).p(C^{1}|C^{2}).p(C^{2}|C^{3}) \ldots p(C^{(h-1)}|C^{(h)})\big) \tag{3}$$

The relation between the conditional probability and the cross-entropy is only valid when the probabilities of the true class are considered i.e. when $y_{true} = 1$.

$$L^{0}_{\text{CE}}(p, C) = -\big[\log p(C^{0}|C^{1}) + \log p(C^{1}|C^{2}) + \ldots + \log p(C^{(h-1)}|C^{(h)})\big] \tag{4}$$

Similarly,

$$L^{1}_{\text{CE}}(p, C) = -\big[\log p(C^{1}|C^{2}) + \log p(C^{2}|C^{3}) \ldots + \log p(C^{(h-1)}|C^{(h)})\big] \tag{5}$$

For any level k, the generalized equation can be written as:

$$\begin{aligned} L^{(k)}_{\text{CE}}(p, C) = -\big[&\log p(C^{(k)}|C^{(k+1)}) + \log p(C^{(k+1)}|C^{(k+2)}) + \\ &\ldots + \log p(C^{(h-1)}|C^{(h)})\big] \end{aligned} \tag{6}$$

$$L^{(k)}_{\text{CE}}(p, C) = -\log p(C^{(k)}|C^{(k+1)}) + L^{(k+1)}_{\text{CE}}(p, C) \tag{7}$$

$$-\log p(C^{(k)}|C^{(k+1)}) = L^{(k)}_{\text{CE}}(p, C) - L^{(k+1)}_{\text{CE}}(p, C) \tag{8}$$

$$\log p(C^{(k)}|C^{(k+1)}) = -\Big[L^{(k)}_{\text{CE}}(p, C) - L^{(k+1)}_{\text{CE}}(p, C)\Big] \tag{9}$$

Hierarchical cross-entropy (HXE) [2] loss is given as:

$$L_{\text{HXE}}(p, C) = -\sum_{l=0}^{h-1} \lambda^{(C^{l})} \log p(C^{(l)}|C^{(l+1)}) \tag{10}$$

$$L_{\mathrm{HXE}}(p, C) = -\big[\lambda^{(C^0)} \log(p^{C^0}|p^{C^1}) + \lambda^{(C^1)} \log(p^{C^1}|p^{C^2}) + \ldots \\ + \lambda^{(C^{(h-1)})} \log(p^{(C^{(h-1)})}|p^{(C^{(h)})})\big] \tag{11}$$

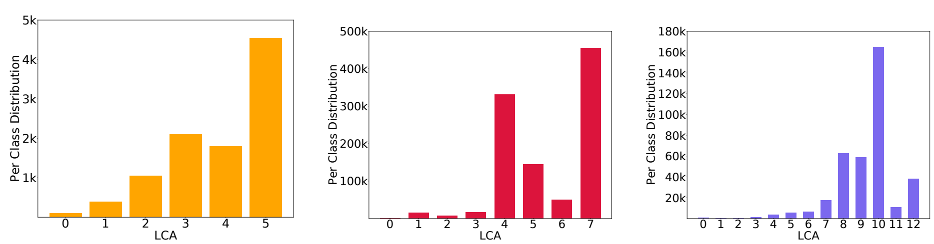
Substituting the value of $\log p(C^{(k)}|C^{(k+1)})$ from Eq. 9

$$L_{\mathrm{HXE}}(p, C) = \lambda^{(C^0)} \left[L^0_{\mathrm{CE}}(p, C) - L^1_{\mathrm{CE}}(p, C)\right] + \lambda^{(C^1)} \left[L^1_{\mathrm{CE}}(p, C) - L^2_{\mathrm{CE}}(p, C)\right] \\ + \ldots + \lambda^{(C^{(h-1)})} \left[L^{(h-1)}_{\mathrm{CE}}(p, C) - L^{(h)}_{\mathrm{CE}}(p, C)\right] \tag{12}$$

$$L_{\mathrm{HXE}}(p, C) = \lambda^{(C^0)} L^0_{\mathrm{CE}}(p, C) + \left[\lambda^{(C^1)} - \lambda^{(C^0)}\right] L^1_{\mathrm{CE}}(p, C) + \\ \left[\lambda^{(C^2)} - \lambda^{(C^1)}\right] L^2_{\mathrm{CE}}(p, C) + \ldots - \lambda^{(C^{(h-1)})} L^{(h)}_{\mathrm{CE}}(p, C) \tag{13}$$

Thus, we have obtained the desired expression where L_{HXE} as weighted sum of cross-entropy loss at different levels of the hierarchy.

2 Analysis of Hierarchical Metrics



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

Fig. 1: Per class distribution of LCA for each dataset.

To plot Figure 1, we first create a symmetric $|\mathcal{C}| \times |\mathcal{C}|$ LCA matrix where $|\mathcal{C}|$ is the total number of fine-grained classes and each entry LCA(i, j) denotes the LCA between class i and class j. For every class, we compute the count of distinct LCA values using this matrix and sum them up for all the classes to plot this distribution for all the datasets. The misclassified samples will likely introduce the errors with LCA value at the peak of the plots. This plot shows the skewness of the hierarchy tree, resulting in larger values of the hierarchical metrics.

Method	LCA sum	Total mistakes
Cross-Entropy	5242.67	2227
YOLO-v2 [6]	11917.33 (-127.31)	3203.33 (-43.84)
HXE α=0.1 [2]	6893.67 (-31.49)	2840.67 (-27.56)
HXE α=0.6 [2]	6965.33 (-32.86)	3041.67 (-36.58)
soft-labels β=4 [2]	7100.33 (-35.43)	3215.33 (-44.38)
soft-labels β=30 [2]	6411 (-22.29)	2699.33 (-21.21)
Chang et al. [3]	5081.33 (3.08)	**2194 (1.48)**
HAF	**4992.67 (4.77)**	2227 (0)
Cross-Entropy + CRM [5]	5128.67 (2.17)	**2223 (0.18)**
HAF + CRM	**4970 (5.02)**	2231.33 (-0.19)

(a) CIFAR-100

Method	LCA sum	Total mistakes
Cross-Entropy	35458.33	14846
YOLO-v2 [6]	43814.33 (-23.57)	18074.33 (-21.75)
HXE α=0.1 [2]	40884.67 (-15.3)	16898 (-13.82)
HXE α=0.6 [2]	41388 (-16.72)	18516 (-24.72)
Soft-labels β=4 [2]	55241(-55.79)	30432 (-104.98)
Soft-labels β=30 [2]	39443 (-11.24)	16976 (-14.35)
Chang et al. [3]	34631.67 (2.33)	15168 (-2.17)
HAF	**33732 (4.54)**	**14859 (0.13)**
Cross-Entropy + CRM [5]	34724 (2.07)	14872.33 (-0.18)
HAF + CRM	**33446 (5.68)**	**14859 (-0.09)**

(b) iNaturalist-19

Table 1: LCA sum i.e. sum of LCA of mistakes and total mistakes on CIFAR-100 and iNaturalist-19. The values reported are the average of three different seeds.

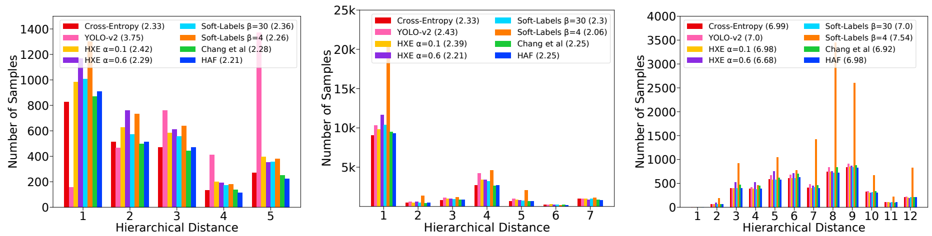
An ideal method is the one that improves the mistakes severity metric while maintaining (or improving) the top-1 accuracy, i.e., the *LCA sum* which is the sum of LCA of misclassified samples, should reduce while maintaining (or improving) the total number of errors. We analyze the *LCA sum* parallel to the total number of mistakes for each of the baseline methods on CIFAR-100 and iNaturalist-19 dataset in the Table 1. The numbers in the parentheses of the column of the *LCA sum* denote the percentage improvement in reducing the LCA sum compared to the baseline cross-entropy. While the numbers in the parentheses of the column of the *Total mistakes* indicate the percentage improvement in reducing the total number of errors compared to the baseline cross-entropy.

3 Mistakes severity using CRM

We plot the distribution of mistakes for methods when evaluated using CRM in Figure 2. Our observations are consistent with Section 5.2 (main text) on all the datasets. CRM benefits most of the methods except Soft-labels $\beta = 4$. The performance of Soft-labels $\beta = 4$ drops when evaluated using CRM. The same reason stated earlier is that the label distribution is flat for smaller β values, leading to predictions with low confidence.

4 Coarse classification Accuracy

In Table 2, 3 and 4 we present the comparisons of HAF with the baselines on coarse-classification accuracy across all hierarchical levels on CIFAR-100, iNaturalist-19 and tieredImageNet-H respectively. L1 refers to Level-1 and is the coarsest-level. We report the coarse-classification accuracy with and without using CRM at test-time. To obtain coarse-classification accuracy, we map the target labels and the predicted labels obtained from the finest-level classifier to their respected coarse classes.We highlight the best entries in each of the column with green . Clearly, CIFAR-100 and iNaturalist-19 surpasses all other methods on both evaluation methods with and without using CRM. On tieredImageNet-H, HAF outperforms other methods towards finer-levels.



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

Fig. 2: Mistakes severity plot showing distributions of mistakes at each level for each dataset when CRM [5] is used for evaluation. Numbers in the bracket denote the mistake severity of the method.

Method	L1	L2	L3	L4
Cross-Entropy	97.12	95.71	90.99	85.94
Barz & Denzler [1]	96.08	94.37	87.53	80.25
YOLO-v2 [6]	95.16	93.05	86.19	78.74
HXE α=0.1 [2]	96.07	94.02	88.17	81.90
HXE α=0.6 [2]	96.47	94.49	88.31	80.82
Soft-labels β=30 [2]	96.39	94.66	89.13	83.39
Soft-labels β=4 [2]	96.19	94.39	88.01	80.67
Chang et al. [3]	97.27	95.86	91.29	86.46
HAF	97.71	96.46	91.81	86.78
Cross-Entropy [5]	97.27	95.94	91.22	86.07
YOLO-v2 [6]	95.24	93.24	86.39	78.86
HXE α=0.1 [2]	96.05	94.04	88.20	81.93
HXE α=0.6 [2]	96.47	94.53	88.42	80.82
Soft-labels β=30 [2]	96.43	94.68	89.10	83.36
Soft-labels β=4 [2]	96.17	94.50	88.28	80.34
Chang et al. [3]	97.47	96.10	91.67	86.68
HAF	97.75	96.59	91.88	86.75

Table 2: Coarse-classification accuracy results on the test set of *CIFAR-100*. The *Top* block reports results without using CRM [5] and the *Bottom* block reports results using CRM on all coarse-levels from level-1(Coarse (L1)) to and level-4(Fine (L4)).

References

1. 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)
2. 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)

Method	L1	L2	L3	L4	L5	L6
Cross-Entropy	97.52	96.99	95.16	88.31	86.24	85.18
Barz & Denzler [1]	97.56	97.03	94.94	85.12	82.68	80.50
YOLO-v2 [6]	97.64	97.08	94.71	84.79	82.09	80.72
HXE α=0.1 [2]	97.41	96.76	94.51	86.06	83.65	82.42
HXE α=0.6 [2]	97.81	97.32	95.30	86.74	84.18	62.89
Soft-labels β=30 [2]	97.55	97.01	95.07	87.22	84.93	83.71
Soft-labels β=4 [2]	97.89	97.40	95.17	85.45	82.77	80.88
Chang et al. [3]	97.75	97.24	95.47	88.91	86.86	85.79
HAF	97.99	97.54	95.86	89.18	87.09	86.01
Cross-Entropy [5]	97.62	97.14	95.43	88.73	86.70	85.60
YOLO-v2 [6]	97.63	97.06	94.70	84.33	81.62	80.17
HXE α=0.1 [2]	97.55	96.92	94.79	86.37	83.98	82.72
HXE α=0.6 [2]	97.83	97.34	95.37	86.97	84.50	83.01
Soft-labels β=30 [2]	97.61	97.06	95.19	87.24	84.95	83.72
Soft-labels β=4 [2]	97.31	96.93	91.80	80.51	77.60	74.30
Chang et al. [3]	97.86	97.37	95.68	89.22	87.19	86.11
HAF	98.02	97.58	95.96	89.38	87.30	86.20

Table 3: Coarse-classification accuracy results on the test set of *iNaturalist-19*. The *Top* block reports results without using CRM [5] and the *Bottom* block reports results using CRM on all coarse-levels from level-1(Coarse (L1)) to and level-6(Fine (L6)).

Method	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11
Cross-Entropy	98.07	97.62	95.17	86.88	79.62	77.40	75.49	73.12	70.86	69.62	69.34
Barz & Denzler [1]	97.87	97.43	94.47	83.73	73.34	70.67	68.03	65.26	62.22	60.52	60.05
YOLO-v2 [6]	98.05	97.55	94.94	85.93	77.84	75.37	73.40	70.84	68.42	67.20	66.91
DeViSE [4]	97.62	97.11	94.36	84.28	75.55	73.07	70.80	67.74	64.97	63.47	63.18
HXE α=0.1 [2]	98.16	97.74	95.32	86.99	79.70	77.61	75.69	73.22	70.96	69.78	69.49
HXE α=0.6 [2]	98.32	97.91	95.44	86.44	78.43	75.92	73.66	70.55	67.57	65.83	65.46
Soft-labels β=30 [2]	98.05	97.59	95.12	86.88	79.76	77.54	75.62	73.25	70.95	69.75	69.45
Soft-labels β=4 [2]	98.01	97.53	94.95	84.85	75.78	73.15	70.69	67.52	64.17	61.56	61.01
Chang et al. [3]	98.10	97.55	95.01	85.87	78.04	75.45	73.34	70.82	68.20	66.76	66.43
HAF	98.11	97.61	95.12	87.01	79.64	77.42	75.55	73.26	71.07	69.80	69.53
Cross-Entropy	98.21	97.82	95.32	87.03	79.68	77.42	75.49	73.10	70.79	69.52	69.24
YOLO-v2 [6]	98.12	97.64	94.97	85.90	77.36	74.89	72.84	70.23	67.78	66.53	66.23
HXE α=0.1 [2]	98.25	97.87	95.46	87.07	79.65	77.51	75.55	73.14	70.85	69.65	69.34
HXE α=0.6 [2]	98.39	97.98	95.53	86.50	78.43	75.92	73.62	70.54	67.44	65.62	65.25
Soft-labels β=30 [2]	98.19	97.78	95.24	87.05	79.78	77.53	75.62	73.30	70.94	69.71	69.40
Soft-labels β=4 [2]	94.57	93.09	80.44	62.97	42.39	33.32	30.49	25.48	21.56	17.43	17.09
Chang et al. [3]	98.24	97.77	95.22	85.97	77.98	75.44	73.35	70.78	68.12	66.64	66.30
HAF	98.25	97.76	95.28	87.12	79.65	77.37	75.48	73.24	70.97	69.74	69.43

Table 4: Coarse-classification accuracy results on the test set of *tieredImageNet-H*. The *Top* block reports results without using CRM [5] and the *Bottom* block reports results using CRM on all coarse-levels from level-1(Coarse (L1)) to and level-11(Fine (L11)).

3. 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)
4. 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
5. 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
6. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263–7271 (2017)