Making Better Mistakes: Leveraging Class Hierarchies with Deep Networks

Luca Bertinetto*

@often_worried
luca.bertinetto@five.ai

Romain Mueller*

@RomainMueller
romain.mueller@five.ai

Konstantinos Tertikas

@ktertikas
konstantinos.tertikas@five.ai

Sina Samangooei

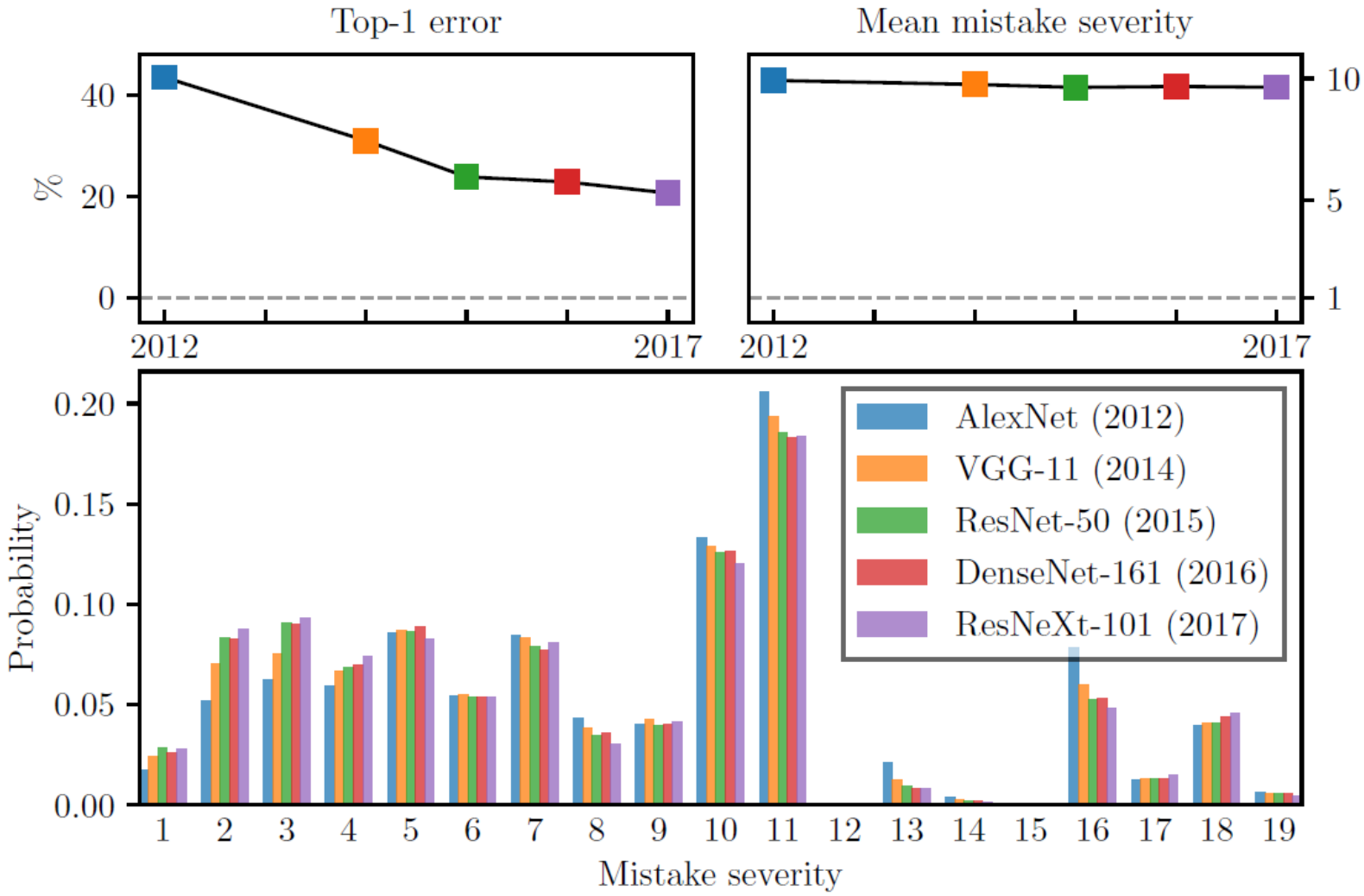
@sinjax
sina@five.ai

Nicholas A. Lord*

@NicholasALord
nick@five.ai

WWW.FIVE.AI

An Observation



Top-1 error and distribution of mistakes w.r.t. the WordNet hierarchy for well-known deep neural network architectures on ImageNet/ILSVRC-12.

Here, a *mistake* is defined as a top-1 prediction which differs from the ground-truth class, and the *severity* of such a mistake is the height of the lowest common ancestor of the predicted and ground-truth classes in the hierarchy.

We see that while the flat top-1 accuracies of state-of-the-art classifiers have improved to impressive levels over the years, the distributions of the severities of the errors that are made have changed very little over this time.

The Issue: Some Mistakes are Worse than Others

?

The relationships between different classes can be important, especially if we might sometimes confuse them.

One natural and convenient way of representing these relationships is through a taxonomic hierarchy tree.

This Idea Is Not New

A survey of hierarchical classification across different application domains

Carlos N. Silla Jr. · Alex A. Freitas

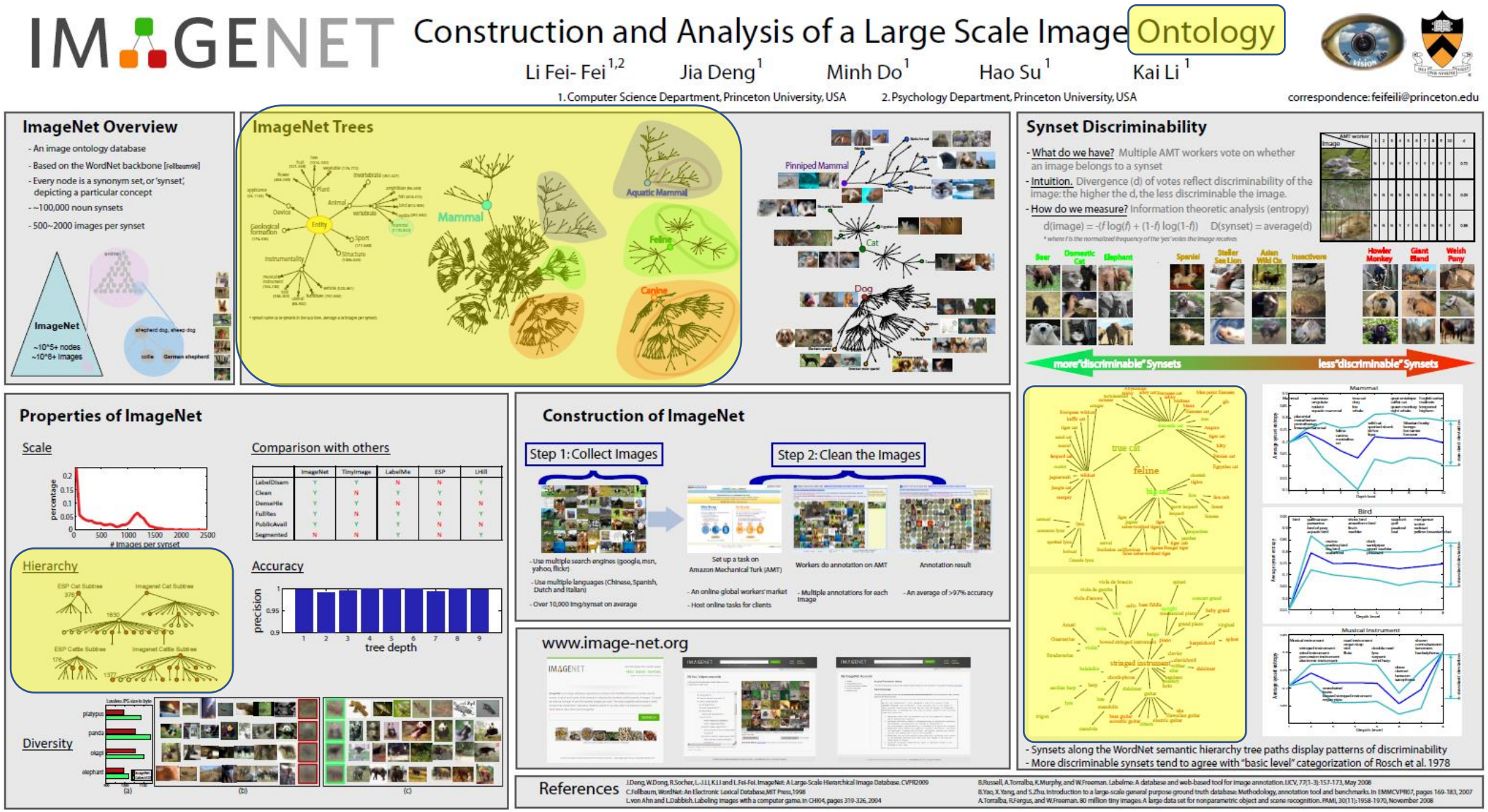
Received: 24 February 2009 / Accepted: 11 March 2010 / Published online: 7 April 2010

Abstract In this survey we discuss the task of hierarchical classification. The literature about this field is scattered across very different application domains and for that reason research in one domain is often done unaware of methods developed in other domains. We define what is the task of hierarchical classification and discuss why some related tasks should not be considered hierarchical classification. We also present a new perspective about some existing hierarchical classification approaches, and based on that perspective we propose a new unifying framework to classify the existing approaches. We also present a review of empirical comparisons of the existing methods reported in the literature as well as a conceptual comparison of those methods at a high level of abstraction, discussing their advantages and disadvantages.

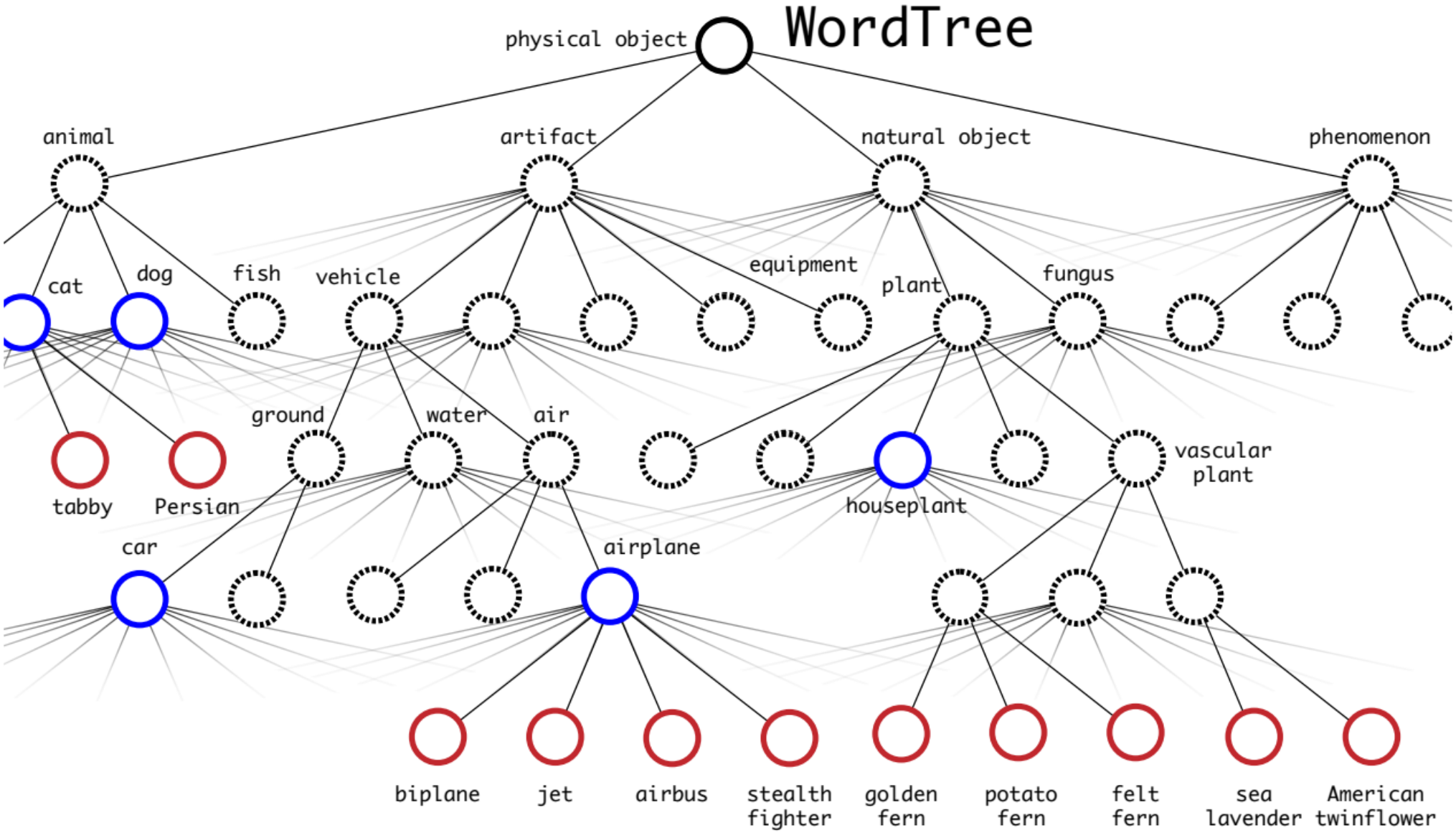
This work assembled and categorised a large collection of hierarchical classification problems and algorithms, and suggested widely applicable measures for quantifying classifier performance in the context of a given hierarchy.

The authors noted that the hierarchy-informed classifiers of the era typically empirically outperformed “flat” (i.e. hierarchy-agnostic) classifiers under standard metrics, with the performance gap increasing further under the suggested hierarchical metrics.

This Idea Is Not New



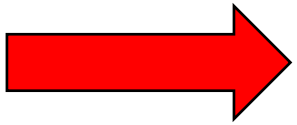
This Idea Is Not New



This Idea Is Not New

ILSVRC 2011

- top-5
- hierarchical



ILSVRC 2012

- top-5
- ~~hierarchical~~

*"In practice we found that all three measures of error (top-5, top-1, and hierarchical) produced the same ordering of results. Thus, since ILSVRC2012 we have been exclusively using the top-5 metric which is the simplest and most suitable to the dataset."**

*Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015.

Our Contributions

We argue that this problem is ripe for revisitation. We...

- Unify a disjoint prior art into a simple framework.
- Propose two methods that are both simple and effective at leveraging class hierarchies to improve performance on hierarchical measures.
  - Each has a single tuning parameter, and each reduces to standard cross-entropy in its respective limit.
- Perform an extensive experimental evaluation against the relevant competitors and point the way to future work.

Framework

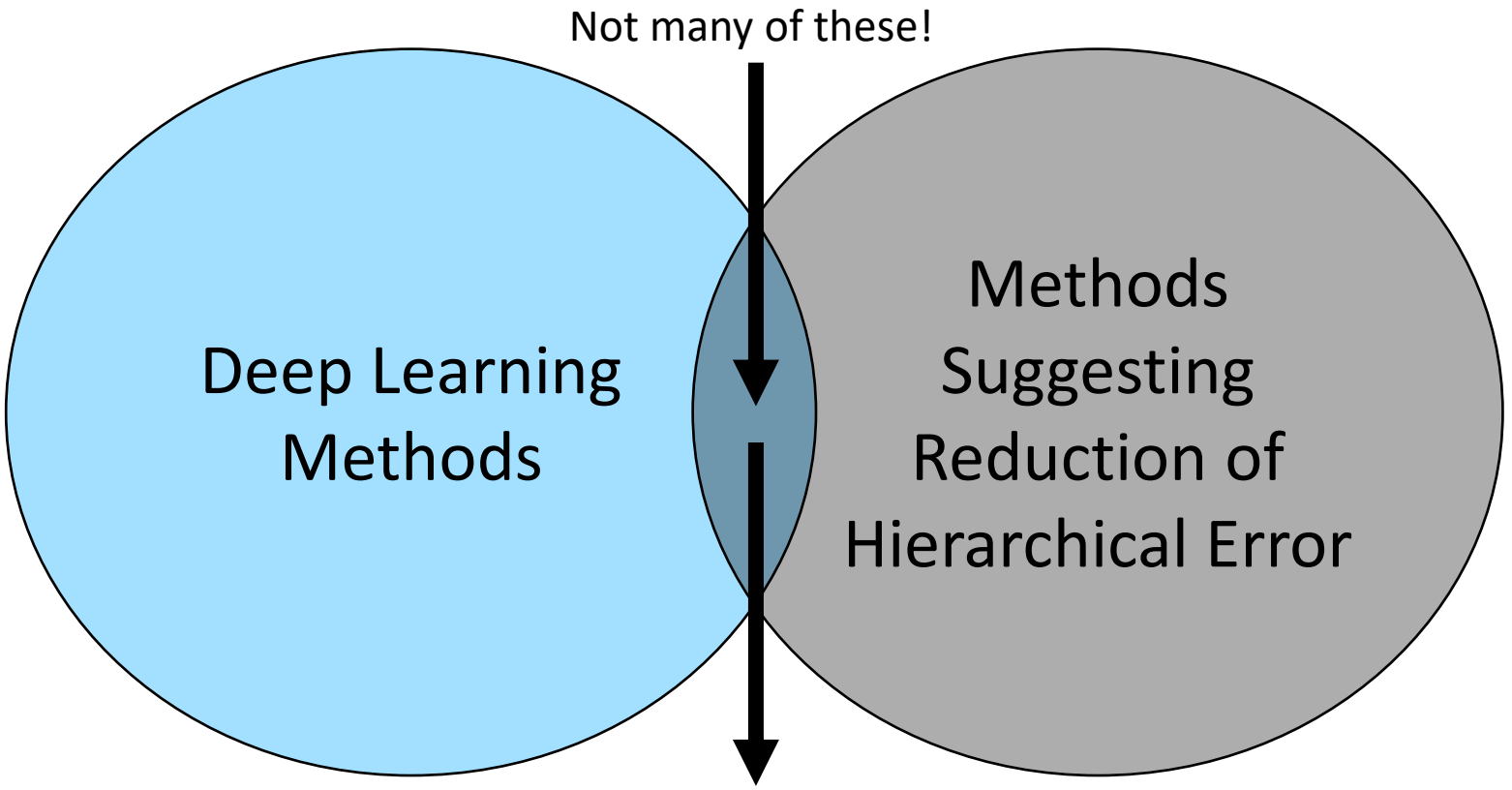
Reconsider the standard supervised learning setup for classification:

$$\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(\phi(x_i;\theta),y(C_i))+\mathcal{R}(\theta)$$

Over a survey of a large number of attempted methods that have to somehow use or represent class relationships, three types emerge:

- Label-embedding methods
  $y(C) \longrightarrow y^{\mathcal{H}}(C)$
- Hierarchical losses
  $\mathcal{L}(\phi(x_i;\theta),y(C_i)) \longrightarrow \mathcal{L}^{\mathcal{H}}(\phi(x;\theta),y(C))$
- Hierarchical architectures
  $\phi(x;\theta) \longrightarrow \phi^{\mathcal{H}}(x;\theta)$
- (well, and one example of $\mathcal{R}^{\mathcal{H}}$)

Framework



- DeViSE (label-embedding)
- Barz & Denzler (label-embedding)
- YOLO9000* (hierarchical architecture)

* (the conditional classifier subsystem, not the main object detection system)

Embarrassingly Simple Method 1: HXE

If hierarchy H is a tree, then it represents a unique factorisation of all categorical probabilities $p(C)$ into the product of the conditional probabilities $p(C^{(l)}|C^{(l+1)})$ along the path from leaf node C to root R:

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

The conditionals can be written in terms of the class probabilities as

$$p(C^{(l)}|C^{(l+1)}) = \frac{\sum_{A \in \text{Leaves}(C^{(l)})} p(A)}{\sum_{B \in \text{Leaves}(C^{(l+1)})} p(B)}$$

where **Leaves**(C) denotes the set of leaves of the subtree rooted at C.

We can then define the hierarchical cross-entropy (HXE) as

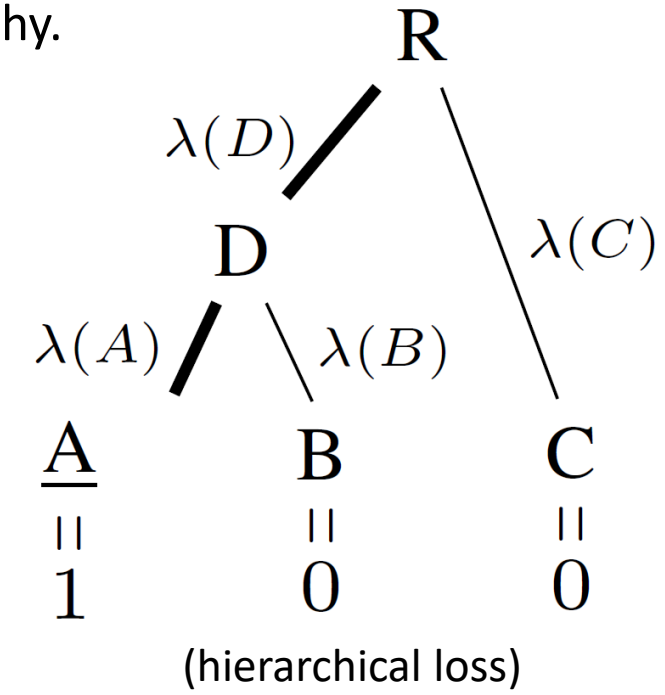
$$\mathcal{L}_{\text{HXE}}(p, C) = -\sum_{l=0}^{h-1} \lambda(C^{(l)}) \log p(C^{(l)}|C^{(l+1)})$$

where $\lambda(C^{(l)})$ is the weight associated with the edge node $C^{(l+1)} \rightarrow C^{(l)}$.

Here, we choose to set

$$\lambda(C) = \exp(-\alpha h(C))$$

where $h(C)$ is the height of node C and $\alpha > 0$ is a hyperparameter that controls the extent to which information is discounted down the hierarchy.



(hierarchical loss)

Embarrassingly Simple Method 2: Soft labels

Label-embedding methods use a mapping function $y(C)$ to associate classes with representations which encode class-relationship information that is absent in the trivial case of the one-hot representation.

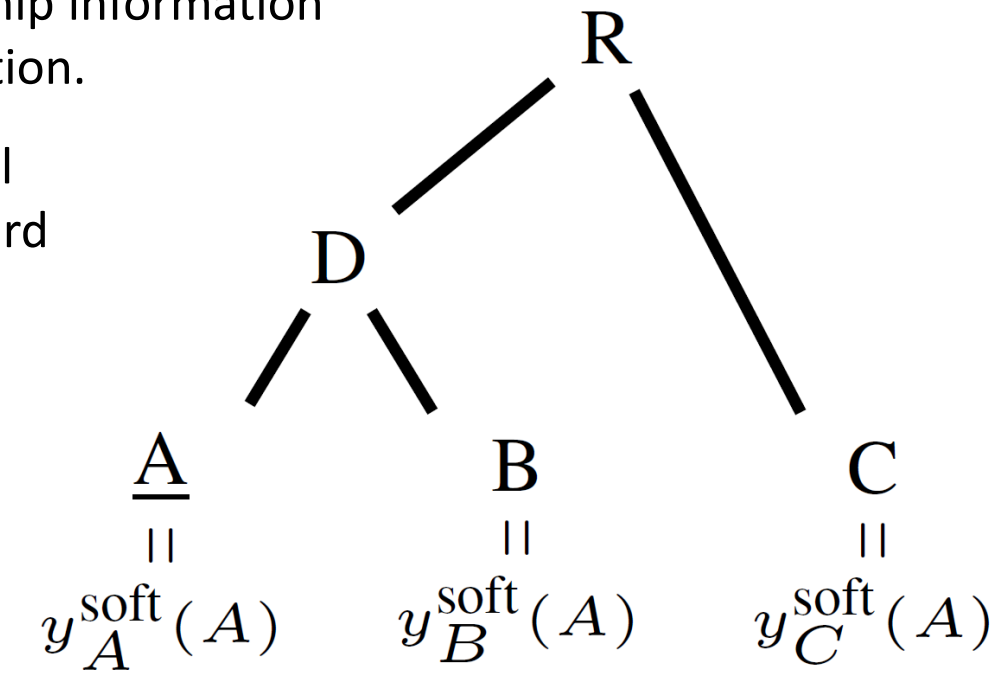
Here, we simply choose a mapping $y^{\text{soft}}(C)$ to a categorical distribution over the classes, allowing us to use the standard cross-entropy loss:

$$\mathcal{L}_{\text{Soft}}(p, C) = -\sum_{A \in \mathcal{C}} y_A^{\text{soft}}(C) \log p(A)$$

where the embedding is given componentwise by

$$y_A^{\text{soft}}(C) = \frac{\exp(-\beta d(A, C))}{\sum_{B \in \mathcal{C}} \exp(-\beta d(B, C))}$$

for class distance function d and parameter β.

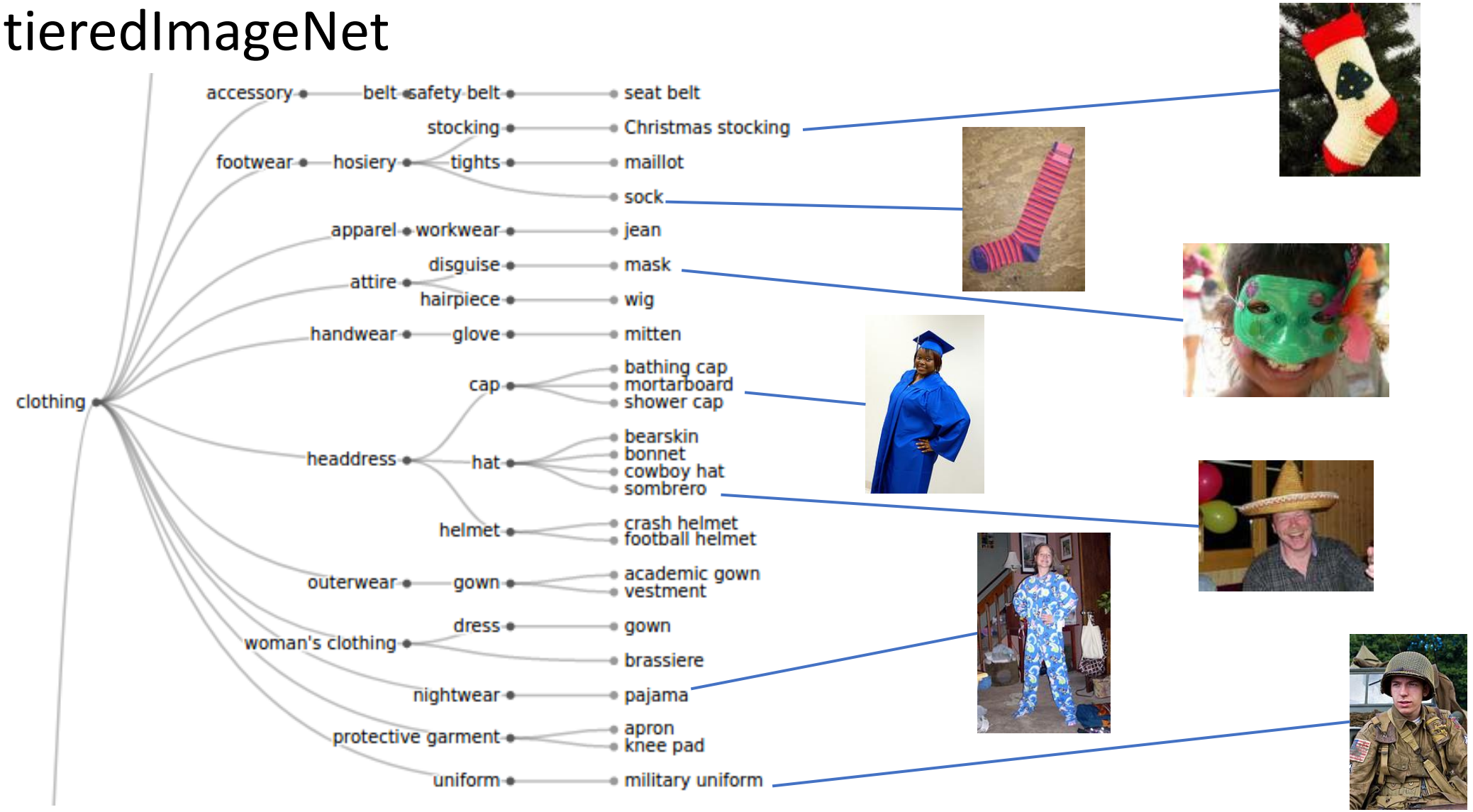


(label embedding)

For the distance function $d(C_i, C_j)$, we use the height of $\mathbf{LCA}(C_i, C_j)$ divided by the height of the tree.

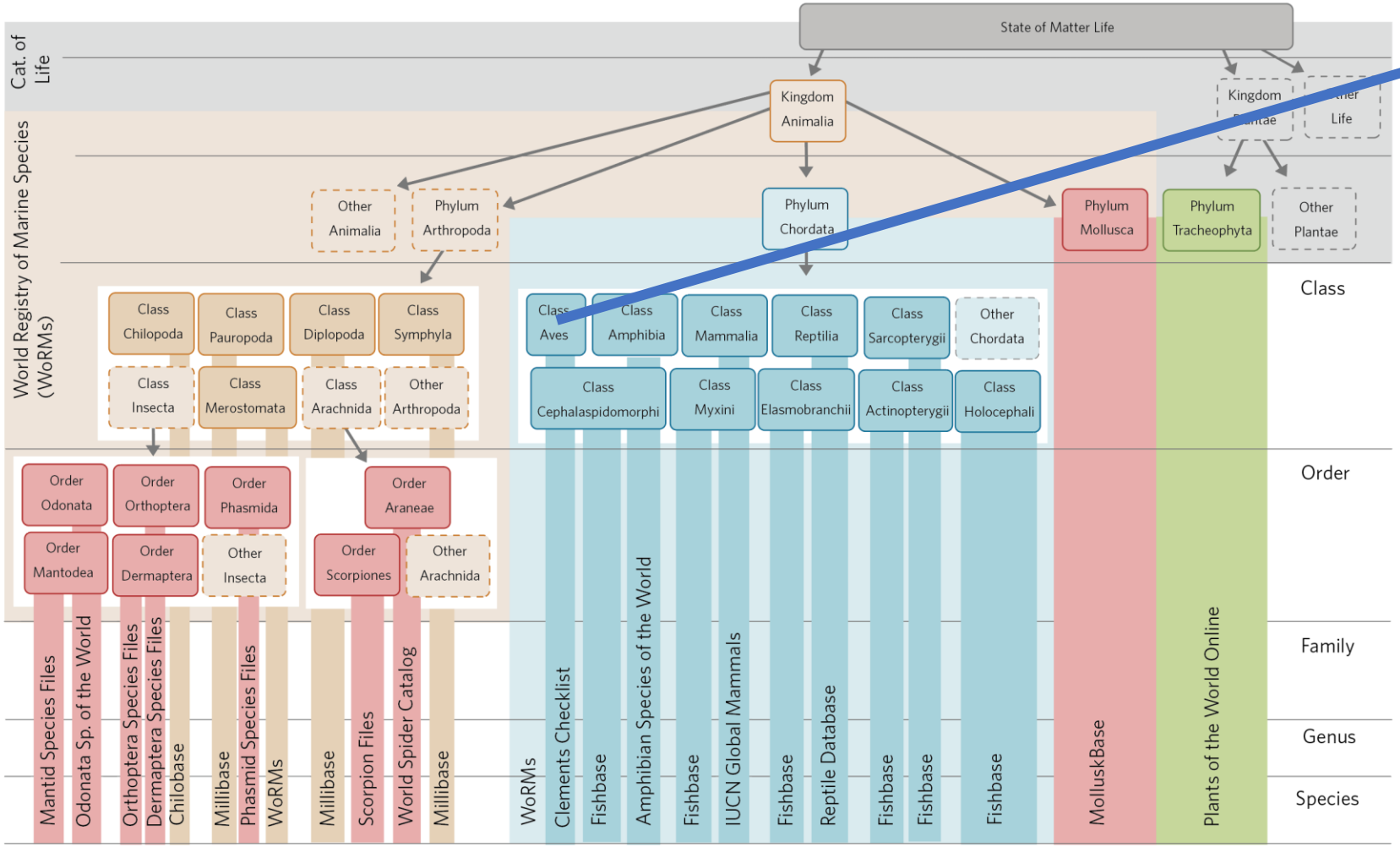
Evaluation: Datasets

tieredImageNet



Evaluation: Datasets

iNaturalist



Birds (Class Aves)

Bald Eagle (*Haliaeetus leucocephalus*) on May 03, 2020

Green Heron (*Butorides virescens*) on May 02, 2020

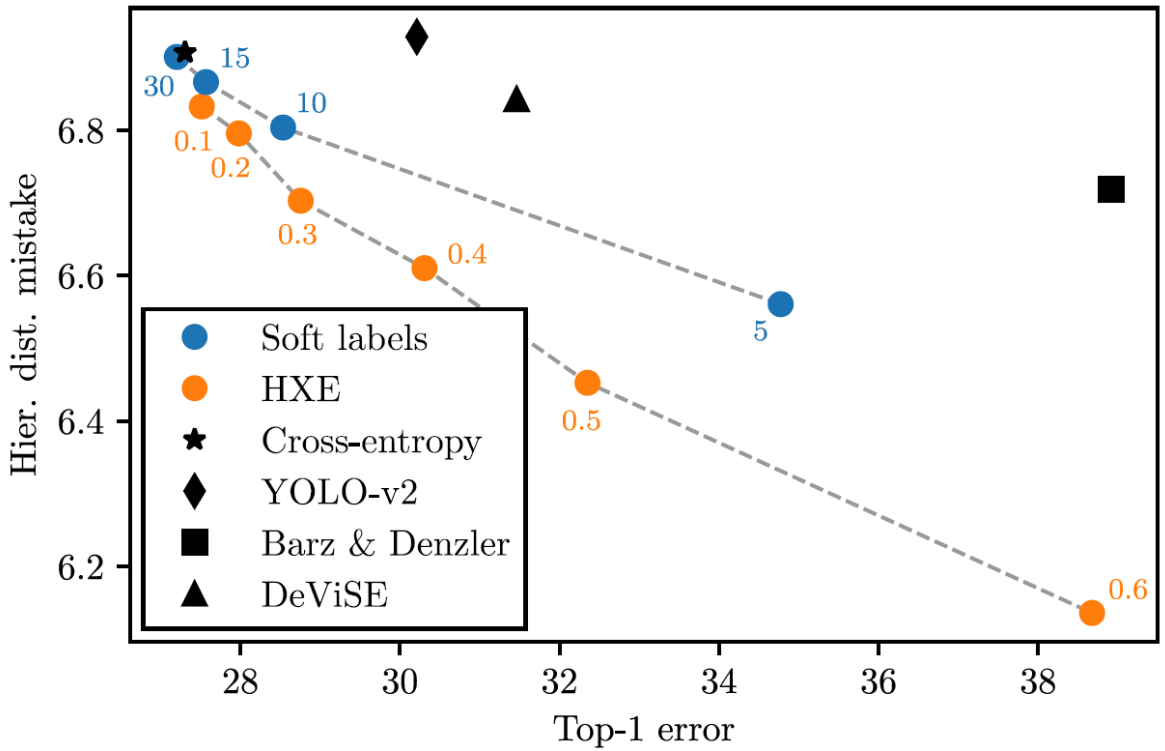
Rose-breasted Grosbeak (*Pheucticus ludovicianus*) on May 06, 2020

American Goldfinch (*Spinus tristis*) on May 06, 2020

Evaluation: Metrics

- "Flat" measure (treats all mistakes equally to one another):
  - **top-k**: the commonly used measure in which an example is considered to be correctly classified if and only if it is among the top k highest-likelihood classes as predicted by the classifier
- Hierarchical measures:
  - **hierarchical distance of a mistake**: the mean height of the LCA between the ground truth and the predicted class *when the input is misclassified*, i.e. when the class with the maximum predicted likelihood is incorrect.
  - **average hierarchical distance of top-k**: mean LCA height between the ground truth and each of the k most likely predicted classes

Evaluation: Results



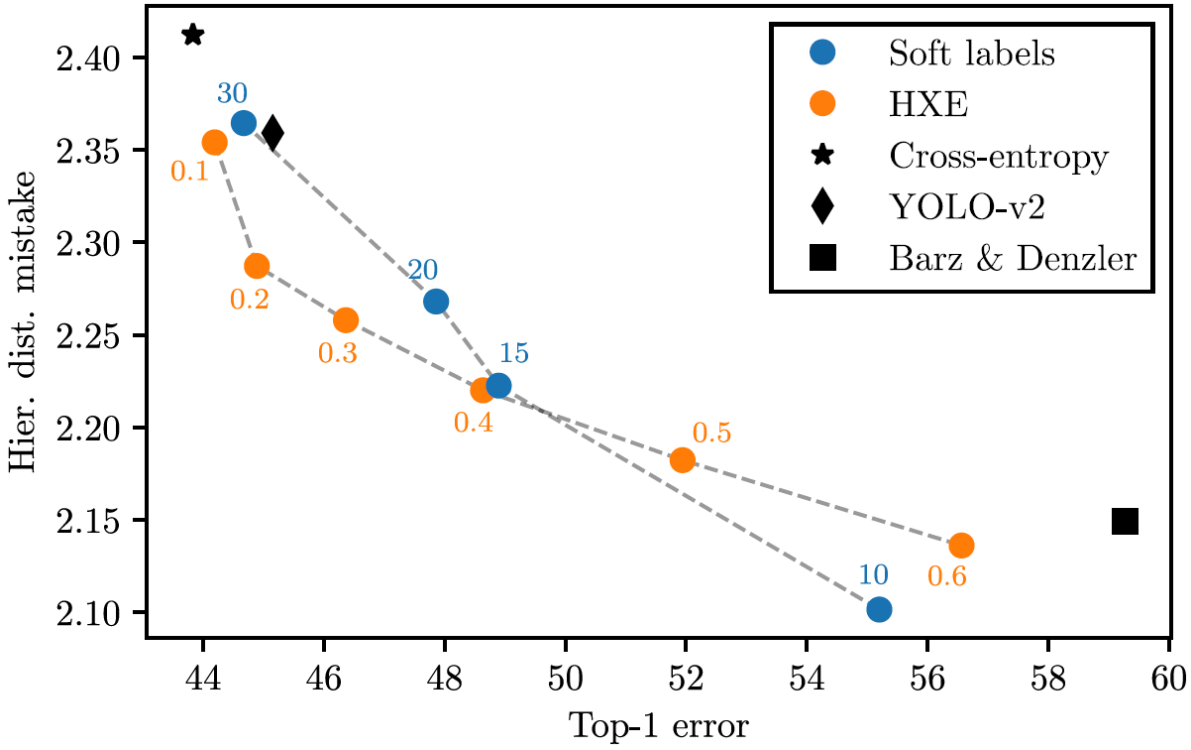


Figure 3: Top-1 error vs. hierarchical distance of mistakes, for *tieredImageNet-H* (top) and *iNaturalist-H* (bottom). Points closer to the bottom-left corner of the plot are the ones achieving the best tradeoff.

Placement towards the lower left of the plots indicates better performance, as it represents reduction of both top-1 and hierarchical error measures.

Evaluation: Results

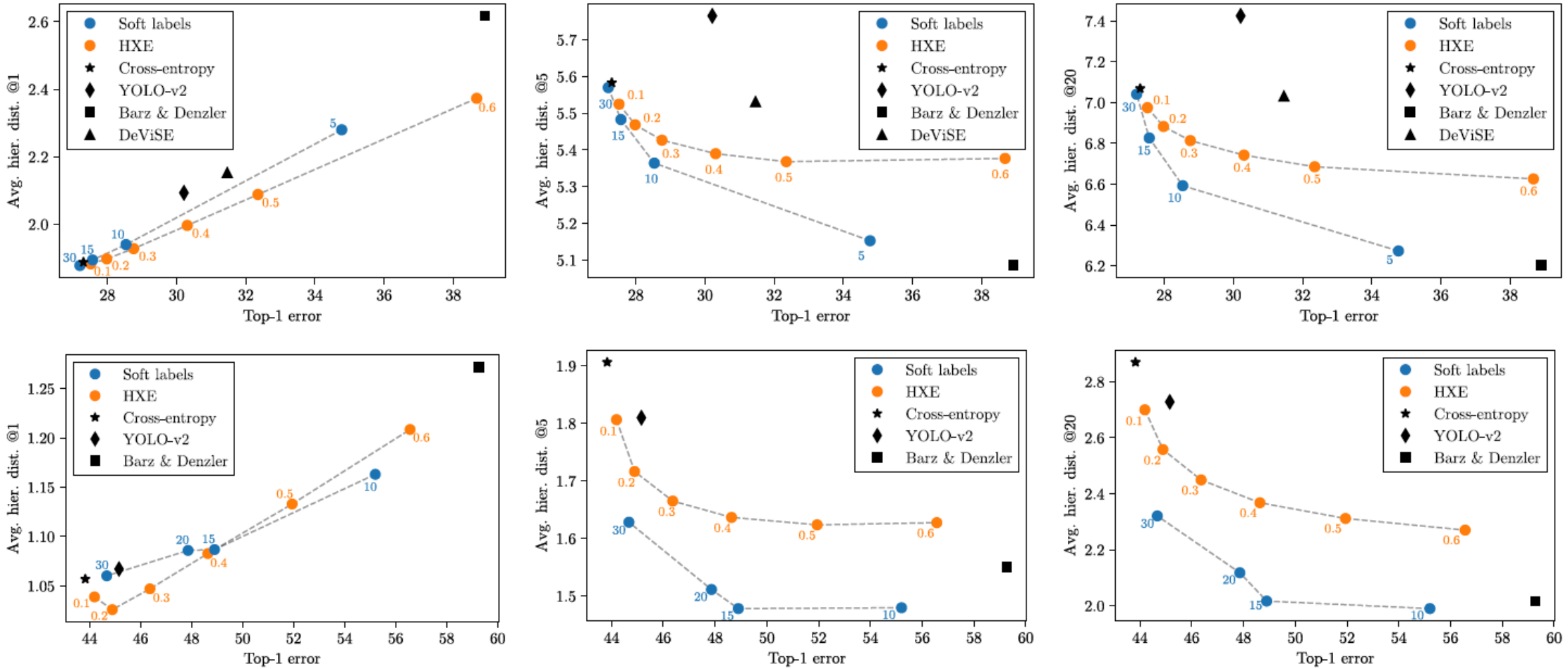


Figure 4: Top-1 error vs. average hierarchical distance of top-k (with $k \in \{1, 5, 20\}$) for *tieredImageNet-H* (top three) and *iNaturalist-H* (bottom three). Points closer to the bottom-left corner of the plot are the ones achieving the best tradeoff.

Evaluation: Results

Table 1: Results on the test sets of *tieredImageNet-H* (top) and *iNaturalist-H* (bottom), with 95% confidence intervals. For each column of each dataset, the best entry is hightlighted in yellow, while the worst is highlighted in gray.

	Hier. dist. mistake ↓	Avg. hier. dist. @1 ↓	Avg. hier. dist. @5 ↓	Avg. hier. dist. @20 ↓	Top-1 error ↓
CROSS-ENTROPY	6.89 ± 0.004	1.90 ± 0.002	5.59 ± 0.004	7.07 ± 0.007	27.55 ± 0.038
BARZ&DENZLER [3]	6.72 ± 0.017	2.62 ± 0.014	5.09 ± 0.009	6.21 ± 0.007	39.03 ± 0.157
YOLO-V2 [27]	6.91 ± 0.006	2.10 ± 0.002	5.77 ± 0.012	7.42 ± 0.018	30.43 ± 0.030
DEVISE [12]	6.83 ± 0.005	2.17 ± 0.003	5.54 ± 0.003	7.04 ± 0.002	31.69 ± 0.058
HXE α=0.1 (ours)	6.83 ± 0.009	1.89 ± 0.003	5.53 ± 0.004	6.98 ± 0.008	27.68 ± 0.066
HXE α=0.5 (ours)	6.46 ± 0.026	2.11 ± 0.021	5.37 ± 0.003	6.69 ± 0.008	32.61 ± 0.443
SOFT-LABELS β=15 (ours)	6.83 ± 0.005	1.90 ± 0.004	5.49 ± 0.002	6.83 ± 0.002	27.78 ± 0.063
SOFT-LABELS β=5 (ours)	6.56 ± 0.009	2.29 ± 0.008	5.16 ± 0.006	6.28 ± 0.005	35.00 ± 0.096
CROSS-ENTROPY	2.41 ± 0.003	1.05 ± 0.004	1.90 ± 0.004	2.87 ± 0.006	43.77 ± 0.138
BARZ&DENZLER [3]	2.19 ± 0.008	1.27 ± 0.007	1.56 ± 0.006	2.03 ± 0.005	57.83 ± 0.137
YOLO-V2 [27]	2.37 ± 0.006	1.07 ± 0.007	1.81 ± 0.008	2.73 ± 0.009	45.23 ± 0.202
HXE α=0.1 (ours)	2.35 ± 0.007	1.04 ± 0.004	1.80 ± 0.004	2.70 ± 0.009	44.28 ± 0.171
HXE α=0.6 (ours)	2.13 ± 0.003	1.21 ± 0.004	1.62 ± 0.003	2.68 ± 0.003	56.61 ± 0.241
SOFT-LABELS β=30 (ours)	2.35 ± 0.002	1.05 ± 0.005	1.62 ± 0.005	2.32 ± 0.004	44.75 ± 0.139
SOFT-LABELS β=10 (ours)	2.10 ± 0.005	1.16 ± 0.006	1.47 ± 0.004	1.99 ± 0.003	55.16 ± 0.196

Evaluation: Results (Random-Hierarchy Experiments)

To investigate the relationship between the nature of the hierarchy itself and the results obtained, we performed the following simple experiment:

1. Randomly shuffled the nodes of the hierarchy tree while preserving the ground-truth labels, thereby destroying any possible correlation between “visual proximity” and “tree proximity”.
2. Repeated the previous experiments using these shuffled hierarchies for both training *and* evaluation.

Evaluation: Results (Random-Hierarchy Experiments)

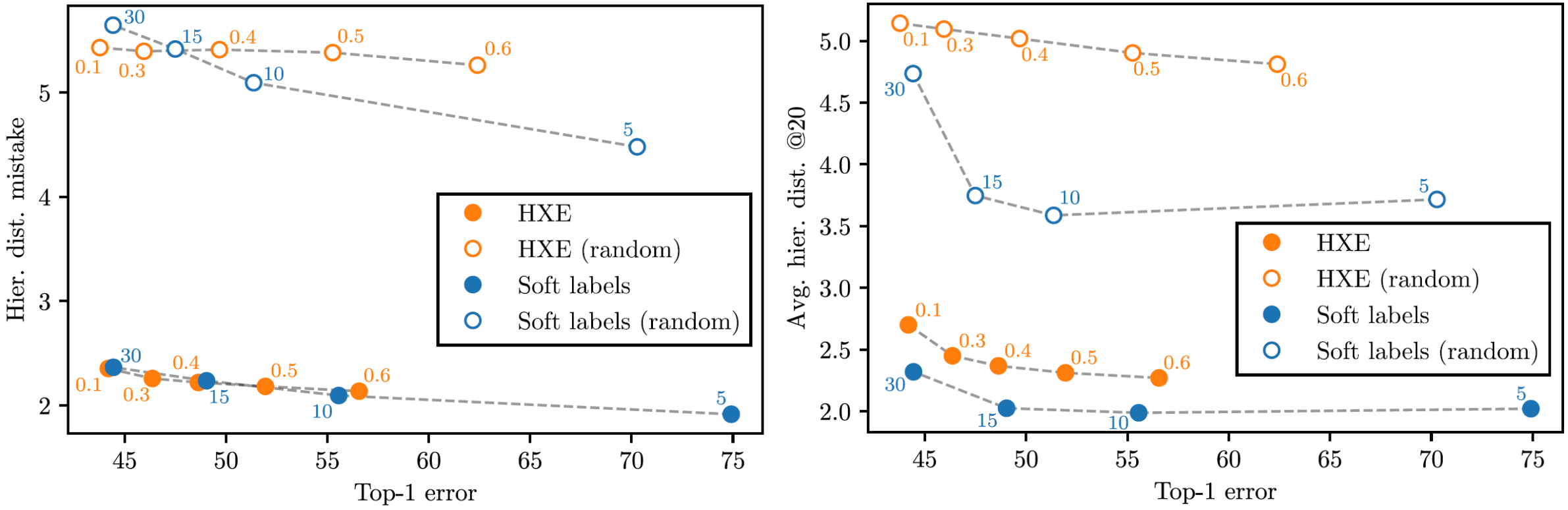


Figure 5: Top-1 error vs. hierarchical distance of mistakes (top) and hierarchical distance of top-20 (bottom) for *iNaturalist-H*. Points closer to the bottom-left corner of the plots are the ones achieving the best tradeoff.

Conclusion

- This problem is still open and deserving of attention.
- Our simple baselines already outperform the few modern methods that have attempted the problem.
- We have seen evidence of an empirical trade-off between top-1 and hierarchical accuracies.
- We have also seen that the hierarchy you use is crucial to the results you're able to obtain.
- We want you to publish a paper beating our results.