MAKING EVERY LABEL COUNT: HANDLING SEMANTIC IMPRECISION BY INTEGRATING DOMAIN KNOWLEDGE

ICPR 2020 Paper #1903

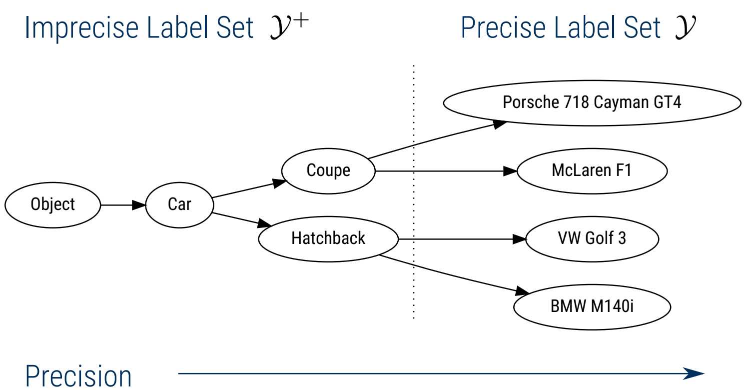
Clemens-Alexander Brust and Björn Barz and Joachim Denzler *Friedrich Schiller University Jena, Germany*

IMPRECISE DATA

Not everyone can label data perfectly. These labels all describe this bird correctly:

- Non-breeding Snow Bunting
- Snow Bunting
- Perching Bird
- Bird
- Object

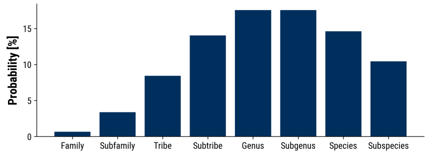
But they are not equally **precise**. We propose to learn from all labels, even **imprecise** ones. This allows us to consider more sources of training data, such as volunteer annotators.



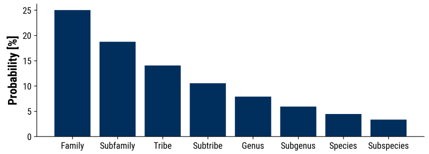
We formally define semantically imprecise data using a class hierarchy. To learn from imprecise data, we adapt a **hierarchical classifier** [1]. Ordinary one-hot encoded softmax classification cannot solve our task because our classes are not all mutually exclusive. Our method **CHILLAX** (Class Hierarchies for Imprecise Label Learning and Annotation eXtrapolation) does not have this limitation.

NOISE MODELS

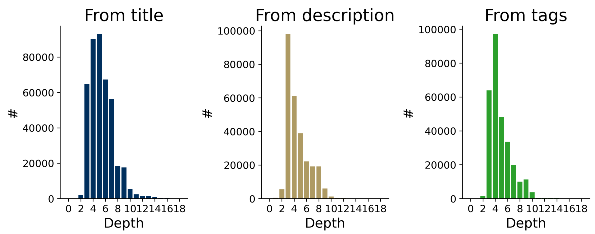
To generate synthetic training data for our experiments, we model **label noise** as a distribution over depth in the class hierarchy.



The figure above shows the Poisson distribution that we expect from **volunteer** labelers who are not experts. Below we show the geometric distribution for data randomly crawled from the **web**.



To validate the correctness of these models, we evaluate the metadata of 1.5M Flickr images uploaded in 2019. Below, we show that distribution resulting from titles, descriptions and tags mapped to WordNet.



EXPERIMENTS

We first investigate the accuracy of our method CHILLAX on the North American Birds dataset with the "**volunteer**" noise model applied to the training data. We compare against two one-hot softmax classifiers as baselines: "leaves only", where all imprecise data is ignored, and "random leaf", where imprecise labels are mapped to a random leaf node in the correct subhierarchy.

Method \ Setting	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	**No Noise**
Baseline: leaves only	26.5 $\pm$ 0.8	61.9 $\pm$ 0.5	74.9 $\pm$ 0.3	79.1 $\pm$ 0.2	82.8 $\pm$ 0.2
Baseline: random leaf	11.1 $\pm$ 0.4	36.8 $\pm$ 0.4	59.0 $\pm$ 0.5	70.6 $\pm$ 0.3	82.8 $\pm$ 0.2
Ours	**42.9** $\pm$ 0.4	**70.1** $\pm$ 0.2	**77.7** $\pm$ 0.3	**80.1** $\pm$ 0.1	81.4 $\pm$ 0.2
Precise samples	4.8	22.7	45.9	65.9	100.0

To test the robustness of our method, we add further label noise by introducing **10% inaccuracy**.

Method \ Setting	$\lambda = 1$	$\lambda = 2$	$\lambda = 3$	$\lambda = 4$	**Only Inacc.**
Baseline: leaves only	22.1 $\pm$ 0.4	54.4 $\pm$ 1.2	67.9 $\pm$ 0.1	**73.1** $\pm$ 0.6	77.3 $\pm$ 0.1
Baseline: random leaf	10.0 $\pm$ 0.3	33.1 $\pm$ 0.6	53.1 $\pm$ 0.7	65.4 $\pm$ 0.2	77.3 $\pm$ 0.1
Ours	**34.6** $\pm$ 1.2	**60.5** $\pm$ 0.3	**69.8** $\pm$ 0.3	72.8 $\pm$ 0.2	75.3 $\pm$ 0.4
Precise samples	4.8	22.7	45.9	65.9	100.0

Finally, we compare against the state-of-the-art **HEX** method [2] on the ImageNet **ILSVRC2012** dataset using their experimental protocol.

Method \ Setting	p=0.99	p=0.95	p=0.9	p=0.5	**No Noise**
HEX	**41.5** (68.5)	**52.4** (77.2)	55.3 (79.4)	58.2 (80.8)	62.6 (84.3)
Ours	38.1 (**68.6**)	52.1 (**78.1**)	**55.5** (**80.2**)	**62.1** (**83.6**)	62.5 (83.5)

[1] Brust, C.-A., & Denzler, J. Integrating domain knowledge: Using hierarchies to improve deep classifiers. Asian Conference on Pattern Recognition (ACPR) 2019.
[2] Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., & Adam, H. Large-Scale Object Classification Using Label Relation Graphs. European Conference on Computer Vision (ECCV) 2014.