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

Clemens-Alexander Brust and Björn Barz and Joachim Denzler

January 15th, 2021

ICPR 2021 Paper 1903

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

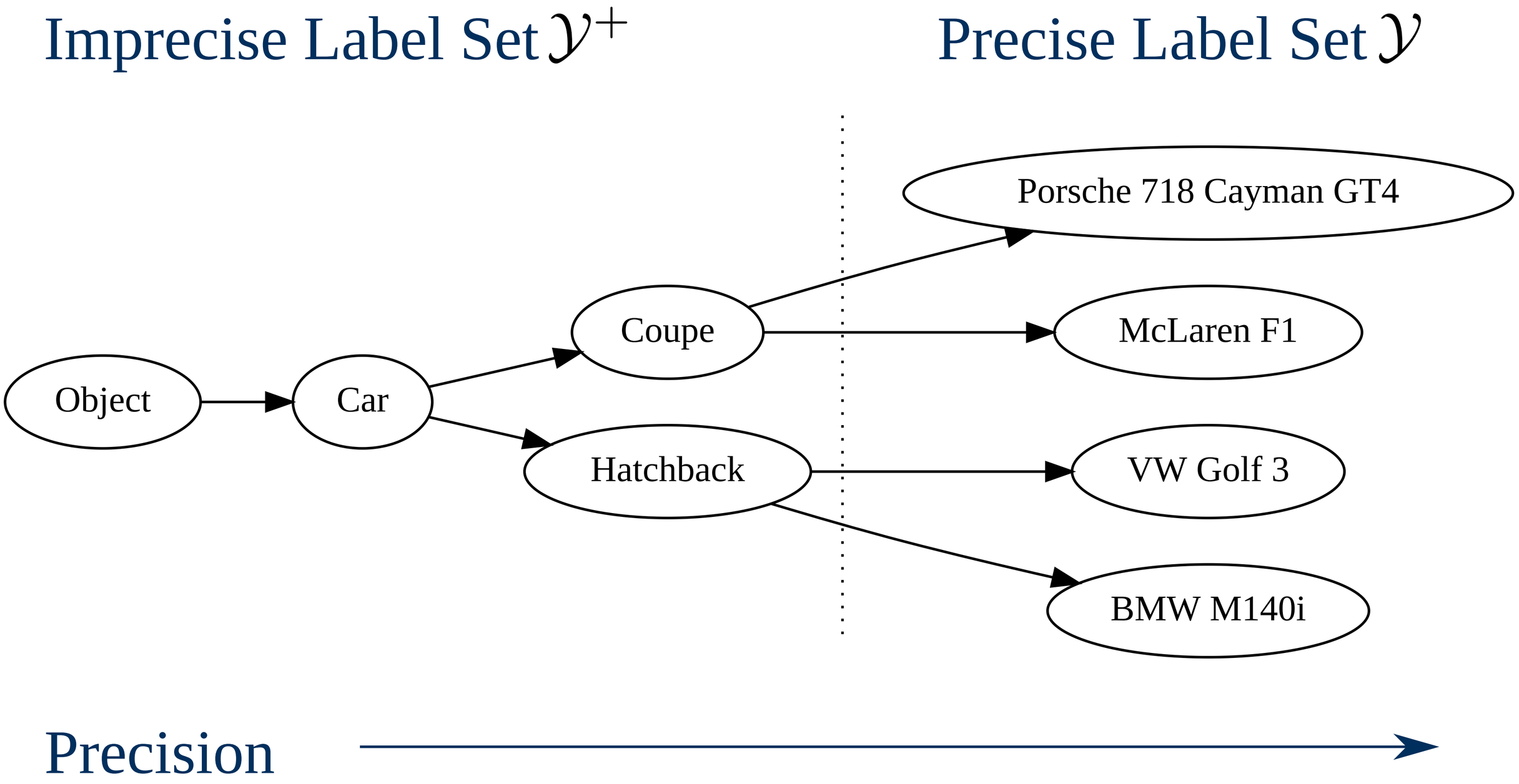
- want 🐦‍🔥 📱
- need $\left(📷, 🏷️ \right)_{i=1,\ldots,n}$
- but 💰 $= 0$

LARGE AMOUNTS OF NOT-SO-WELL-LABELED DATA

Cheap data, but noisy in certain areas:

- **Accuracy**: a *Yellow-billed Magpie* is labeled as *Black-billed Magpie*, *Barn Owl*, or *Ferrari 488 Pista*.
- **Precision**: a *Snow Bunting (Nonbreeding)* is labeled as *Snow Bunting*, *Perching Bird*, *Bird* or *Object*.

Literature focuses on accuracy, rarely even considers semantics.
Today: focus on lack of (semantic) precision.


Imprecise Label Set $\mathcal{Y}^+$
Precise Label Set $\mathcal{Y}$
Object
Car
Coupe
Hatchback
Porsche 718 Cayman GT4
McLaren F1
VW Golf 3
BMW M140i
Precision

Task:

- learn from $\mathcal{Y} \cup \mathcal{Y}^+$
- always predict $\mathcal{Y}$

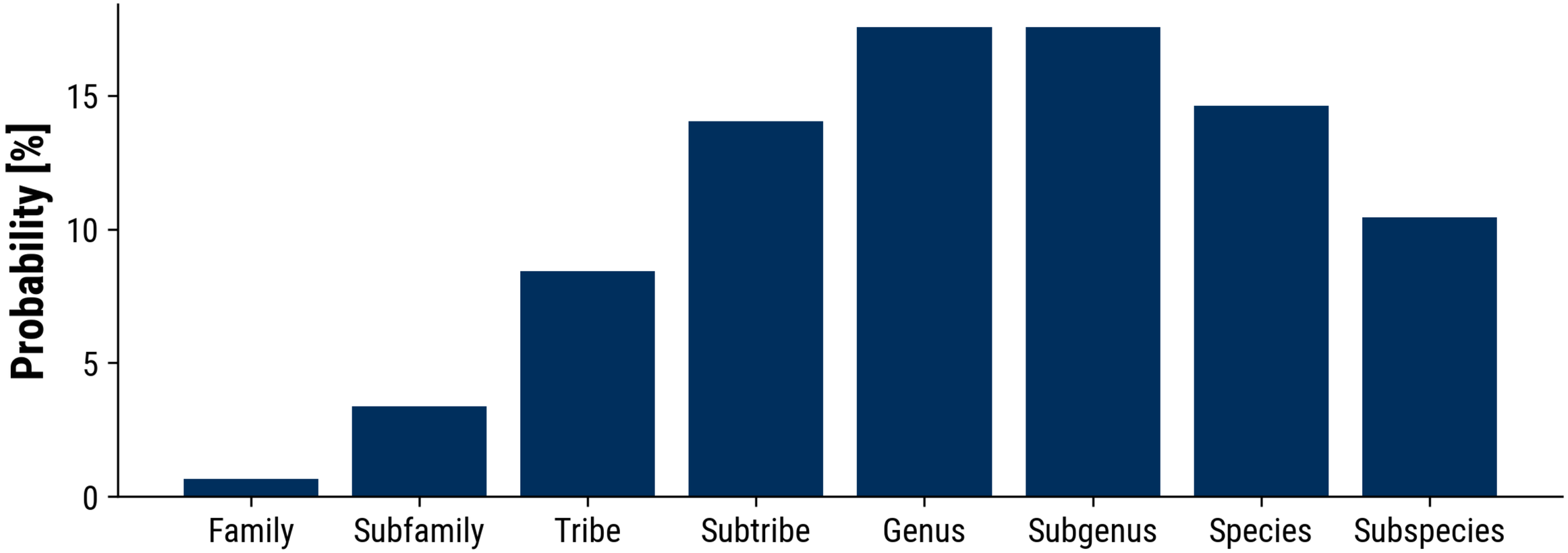
CHILLAX: MODIFIED HIERARCHICAL CLASSIFIER

Typical classifier: model $P(Y|X)$, *e.g.,* using a large neural network, with $\Omega_Y = \mathcal{Y}$.

Our hierarchical classifier [Brust ACPR'19]:

- Model $P\big(Y|\mathrm{parents}(Y) = 1, X\big)$ using a neural network, with $\Omega_Y = \mathcal{Y} \cup \mathcal{Y}^+$
- Learn conditionally using a loss mask. "Learn only what we know."
- To calculate $P(Y|X)$, evaluate $P\big(Y|\mathrm{parents}(Y), X\big)$ recursively up to root.
- Compute argmax over all "allowed" Y, depending on task:
  - only leaf nodes (annotation extrapolation), or
  - all nodes, *e.g.,* to model unconfident predictions.

NOISE MODEL: VOLUNTEERS 🙋



Poisson distribution with $\lambda = 5$.

EXPERIMENTAL RESULTS: VOLUNTEERS 🙋

Noise model: Poisson distribution. *No Inaccuracy*.

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

Table shows accuracy on NABirds validation set (%). 6 runs each.

EXPERIMENTAL RESULTS: VOLUNTEERS 🙋

Noise model: Poisson distribution. *Inaccuracy: 10%*.

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

Table shows accuracy on NABirds validation set (%). 6 runs each.

COMPARISON TO DENG ET AL. ECCV 2014:

- Deng et al. 2014: "Large-Scale Object Classification Using Label Relation Graphs".
- CRF that describes the relationships between concepts:
  - Subsumption and
  - Exclusion.
- Have to perform exact inference for each prediction $\longrightarrow$ very expensive.
- Experiment on ILSVRC2012 classification dataset.
  - Relabel a fraction of samples to their immediate parents.

COMPARISON TO DENG ET AL. ECCV 2014 (2):

Noise model: Relabeling to immediate parents, cf. [Deng ECCV'14]. *No Inaccuracy*.

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)
CHILLAX (Ours)	38.1 **(68.6)**	52.1 **(78.1)**	**55.5 (80.2)**	**62.1 (83.6)**	62.5 (83.5)

Table shows top-1 (top-5) accuracy on ILSVRC2012 validation set (%).

CONCLUSION

Our noise models:

- Capture various sources of labels: volunteers, web crawling...
- Are validated by real-world observations (see paper)

Our method is:

- Somewhat robust to inaccuracy on top of imprecision.
- Competitive w.r.t. [Deng ECCV'14], but not always better.

$\longrightarrow$ Don't throw away imprecise labels!

Thanks!