To Each Metric Its Decoding: Post-Hoc Optimal Decision Rules of Probabilistic Hierarchical Classifiers

Roman Plaud [1,2] Alexandre Perez-Lebel [3,4] Matthieu Labeau [1]
Antoine Saillenfest [2] Thomas Bonald [1]

[1]Institut Polytechnique de Paris [2]Onepoint [3]Inria Saclay [4]Fundamental Technologies, USA

TL;DR

- We formalize how to **optimally make a prediction** from outputs of a hierarchical classifier, with respect to a specified metric.
- For *single-node* predictions, we propose universal metric-optimal algorithms.
- For *subset of nodes* predictions, we derive optimal rules specifically for hierarchical F_β scores.
- Our methods consistently **outperform standard heuristics methods**, particularly in ambiguous or underdetermined cases.

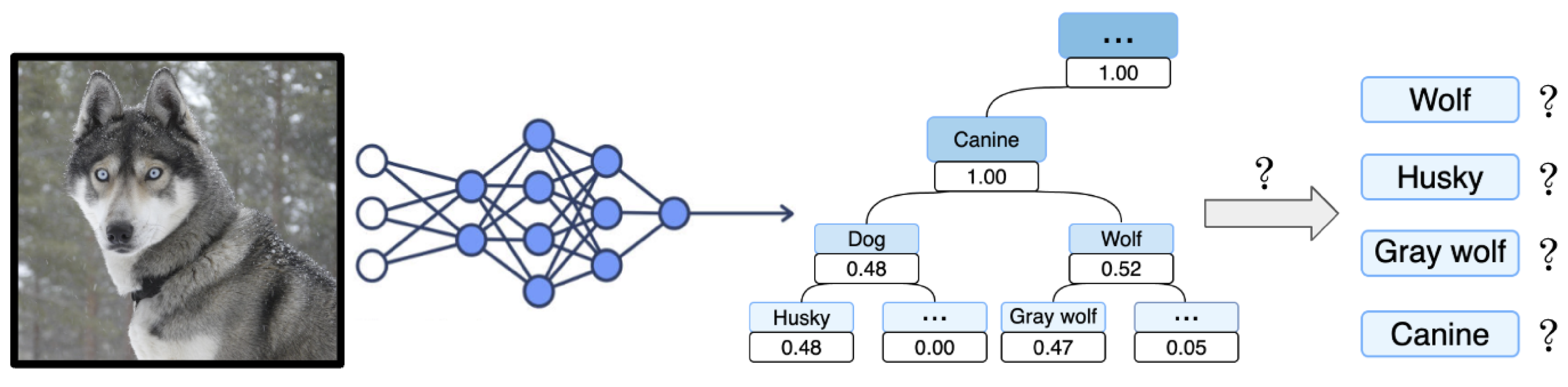
Problem Setup

Given:

- Input x (image, text etc.)
- Model $f \Rightarrow \hat{p}(\cdot \mid X = x)$
- Cost function $C(h, y)$

Objective:

- Find prediction h that is optimal for metric C and for probability estimates $\hat{p}(\cdot \mid X = x)$



Hierarchical Classification

Single leaf Classification:

- Input $x \in \mathcal{X}$, label $y \in \{l_1, \dots, l_K\}$
- Joint distribution $(x, y) \sim \mathbb{P}$

Hierarchy:

- A directed tree $T = (\mathcal{N}, \mathcal{E})$ with leaves $\mathcal{L} = \{l_1, \dots, l_K\}$
- Internal nodes represent super−categories

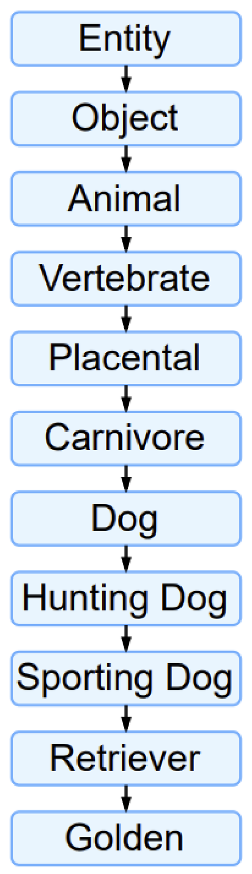


Image of a Golden retriever (top), annotated with its labels in the ImageNet hierarchy (right)

Different metric settings

Evaluation metric. Given prediction set $\mathcal{H}$ and leaf labels $\mathcal{L}$, define

$$C : \mathcal{H} \times \mathcal{L} \to \mathbb{R}$$
$$(h, y) \mapsto C(h, y)$$

Leaf prediction: $\mathcal{H} = \mathcal{L}$

Node prediction: $\mathcal{H} = \mathcal{N}$

Subset of nodes prediction: $\mathcal{H} = \mathcal{P}(\mathcal{N})$

Bayes-optimal decoding

Optimal decision rule. An optimal decision rule for metric $C : \mathcal{H} \times \mathcal{L} \to \mathbb{R}$ is given by $\xi^*_C : \Delta(\mathcal{L}) \to \mathcal{H}$ where

$$\xi^*_C(p) = \underset{h \in \mathcal{H}}{\mathrm{argmin}} \sum_{l \in \mathcal{L}} p(l) C(h, l)$$

Brute-force Decoding: Enumerates all possible predictions. Time complexity: $\mathcal{O}(|\mathcal{H}| \cdot |\mathcal{L}|)$

Objective: Find optimal algorithms with better complexity.

Theoretical Contributions

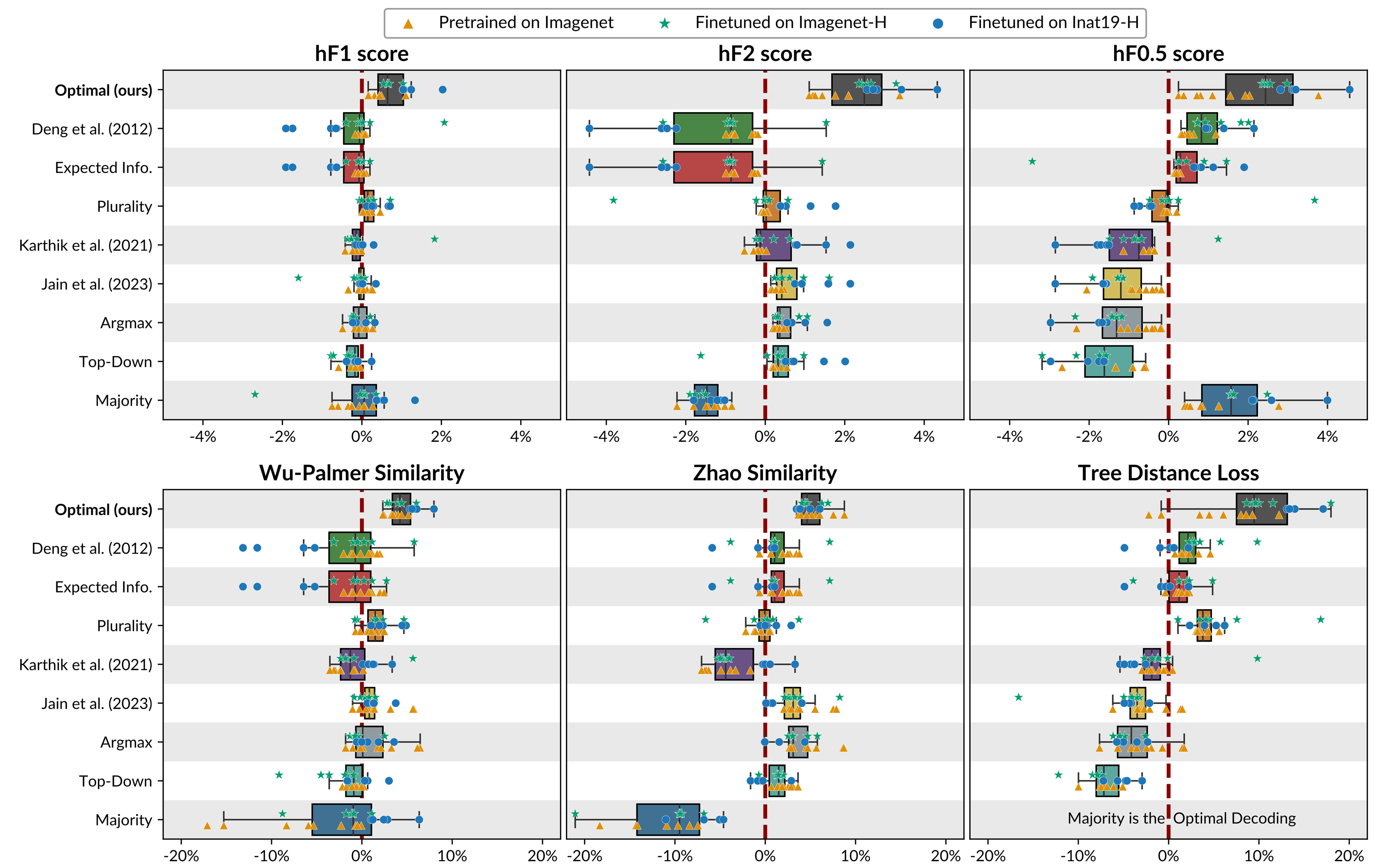
$\mathcal{H}$	Assumption	Brute Force	Our Algorithm	In the paper
$\mathcal{N}$	Hierarchically reasonable	$O(\lvert\mathcal{N}\rvert \times \lvert\mathcal{L}\rvert)$	$O(\log(\lvert\mathcal{N}\rvert) \times \lvert\mathcal{L}\rvert)$	*Theorem 4.4*
$\mathcal{P}(\mathcal{N})$	hF_β scores	$O(2^{\lvert\mathcal{N}\rvert} \times \lvert\mathcal{L}\rvert)$	$O(\log(\lvert\mathcal{N}\rvert)^2 \times \lvert\mathcal{L}\rvert)$	*Theorem 4.7*

Hierarchically Reasonable: C is an increasing function of the length of the shortest path between node h and leaf y. (*Definition 4.2*)

hF_β score: Extension to hierarchical classification of standard F_β-score: balances precision and recall (*Kosmopolous et al., 2014*)

Paper Code Roman Plaud

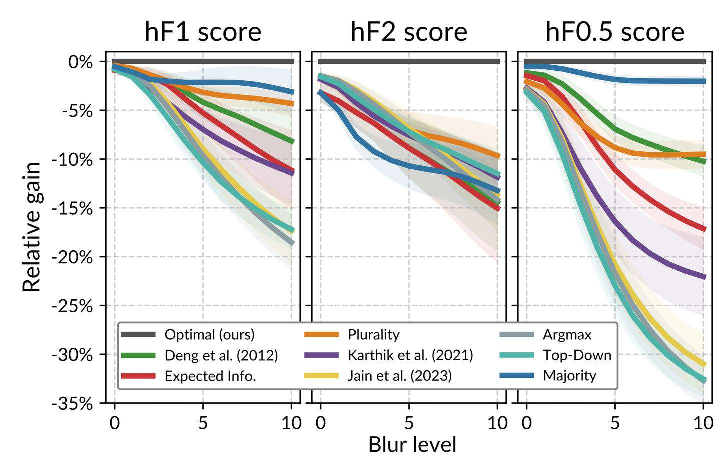
Empirical Results



Relative gain of performance of a decoding strategy vs. the average of all decoding strategies for different metrics.

On the influence of blurring

More model **entropy**
⇒ more heuristic/optimal **disagreements**
⇒ optimal algorithms **crucial**.



Take Home Message

- Our decoding algorithms are **faster** than brute-force decoding and **better** than heuristic decodings.
- The more uncertain a model is, the more important it becomes to optimally decode its outputs.

contact : roman.plaud@telecom-paris.fr