Uncertainty Fingerprints

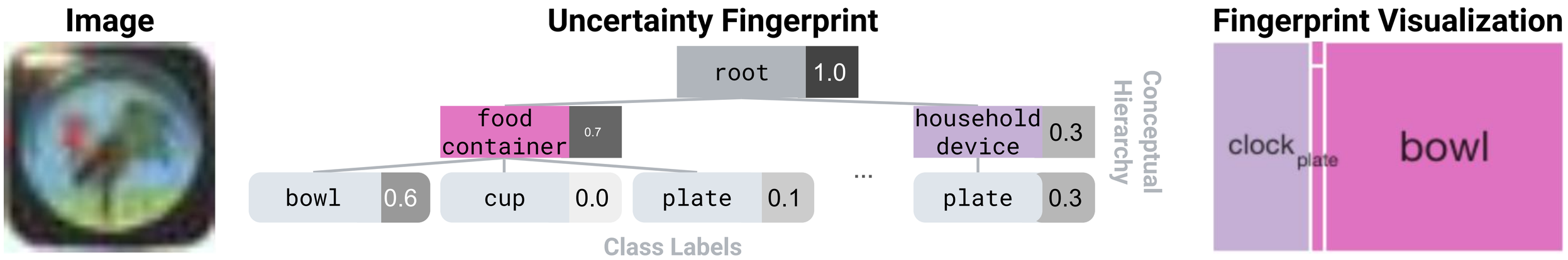
Interpreting Model Decisions with Human Conceptual Hierarchies

Angie Boggust
aboggust@csail.mit.edu
@angie_boggust

Hendrik Strobelt
hendrik@strobelt.com
@hen_str

Arvind Satyanarayan
arvindsatya@mit.edu
@arvindsatya1

Uncertainty Fingerprints describe a model's decision making process. By propagating model output probabilities through a human conceptual hierarchy, they describe the model's confidence in *many* concepts, ranging from low-level classes to abstract ideas. Fingerprint visualizations concisely represent this hierarchy of confidences for glanceable human interpretation.



Interpretability Workflows By computing the entropy at each level of the hierarchy, we can quantify the similarity between fingerprints. In interpretability workflows, comparing uncertainty fingerprints can help users characterize model decisions, identify patterns of confusion, and rethink existing human hierarchies.

Uncertainty Fingerprints: Interpreting Model Decisions with Human Conceptual Hierarchies
Angie Boggust, Hendrik Strobelt, and Arvind Satyanarayan
AI & HCI Workshop at the International Conference on Machine Learning (ICML), Honolulu, USA