



Intelligent
Embedded Systems

dopanim: A Dataset of Doppelganger Animals with Noisy Annotations from Multiple Humans

Marek Herde, Denis Huseljic, Lukas Rauch, and Bernhard Sick
`marek.herde@uni-kassel.de`

Intelligent Embedded Systems, University of Kassel, Germany

November 29, 2024

Motivation

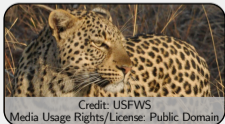
Human Annotators

Objective: Collect a dataset for research purposes containing different data types that can be collected during annotation campaigns with error-prone, human annotators (e.g., crowdworkers).

Task Data

Task:

Classify the animal shown in the image.



Annotation Data



Metadata:
Medium Interest
in Zoology

| | |
|---------|-----|
| Jaguar | 50% |
| Leopard | 20% |
| Cheetah | 30% |

*Annotation
Time:*



N/A



Metadata:
Low Interest
in Zoology

| | |
|---------|-----|
| Jaguar | 30% |
| Leopard | 10% |
| Cheetah | 60% |

*Annotation
Time:*



N/A



Metadata:
High Interest
in Zoology

| | |
|---------|-----|
| Jaguar | 70% |
| Leopard | 30% |
| Cheetah | 0% |

*Annotation
Time:*

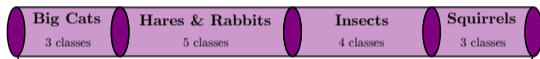
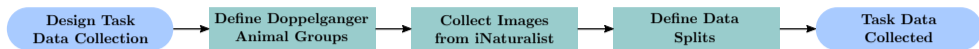


| | |
|---------|------|
| Jaguar | 100% |
| Leopard | 0% |
| Cheetah | 0% |

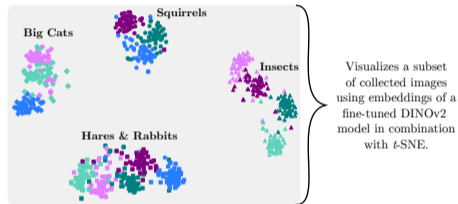
*Annotation
Time:*



Task Data



There is a high similarity among animal classes within each group.

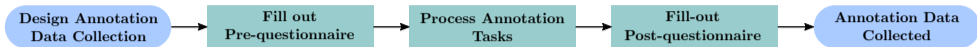


There is no overlap in photographers between train, test, and validation splits.

| Dataset | dopanim |
|--------------------------|-----------|
| | Task Data |
| data modality | image |
| training instances [#] | 10,484 |
| validation instances [#] | 750 |
| test instances [#] | 4,500 |
| classes [#] | 15 |

dopanim: A Dataset of Doppelganger Animals

Annotation Data



Self-assessment Questions

How would you rate your interest in animals and wildlife?

Very low
 Below average
 Average
 Above average
 Very high

How would you rate your knowledge about animals and wildlife?

Very low
 Below average
 Average
 Above average
 Very high

...

[Submit](#)

Self-assessment Questions: General


How do you estimate the average accuracy [%] of the class label to which you assigned the highest likelihood?

How do you rate the average quality of your assigned label likelihoods?

Very low
 Below average
 Average
 Above average
 Very high

...

[Submit](#)



Big Cats

Jaguar

Leopard

Cheetah

Label Likelihoods

Squirrels

Douglas' Squirrels

American Red Squirrel

Eurasian Red Squirrel

Label Likelihoods

Hares & Rabbits

Brown Hare

Jackrabbit

Marsh Rabbit

European Rabbit

Desert Cottontail

Label Likelihoods

Insects

Asian Hornet

European Hornet

European Paper Wasp

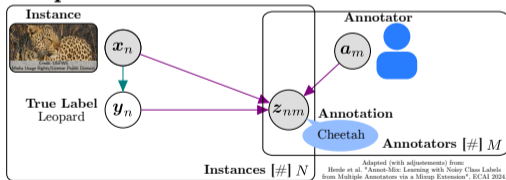
German Yellowjacket

Label Likelihoods

| Dataset | dopanim |
|--|-------------------------------------|
| Annotation Data | |
| annotators [#] | 20 |
| annotation platform | LabelStudio |
| annotator meta-data | <input checked="" type="checkbox"/> |
| annotation times | <input checked="" type="checkbox"/> |
| soft class labels | <input checked="" type="checkbox"/> |
| annotations per instance [$\bar{\#}$] | 5.0 \pm 0.19 |
| annotations per annotator [$\bar{\#}$] | 2,602 \pm 1,255 |
| overall accuracy [%] | 67.3 |
| accuracy per annotator [%] | 65.6 \pm 14.7 |

Multi-annotator learning approaches consider which class label originates from which annotator to estimate the **annotators' performances** (e.g., **confusion matrices**) for improving **neural networks' generalization performances** during training.

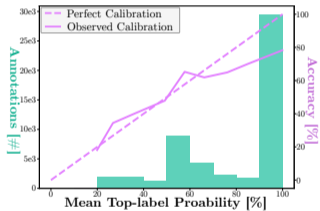
Graphical Model:



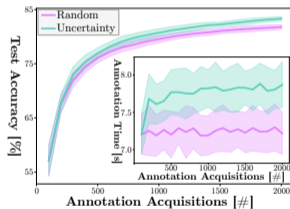
Benchmark: The empirical evaluation covers

- 7 dataset variants of **dopanim** with varying noise rates and numbers of annotations per instance,
- 9 multi-annotator learning approaches with different assumptions regarding annotators' performances,
- 3 evaluation scores in the form of accuracy, Brier score, and top-calibration error.

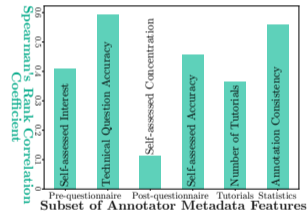
Beyond Hard Class Labels



Active Learning with Real Annotation Times



Learning from Annotator Metadata



Takeaway: dopanim is a multi-purpose image classification dataset supporting research in many areas, e.g., noisy label learning, active learning, and learning beyond hard class labels.

Dataset @ Zenodo

<https://zenodo.org/records/14016659>



`marek.herde@uni-kassel.de`

Code @ GitHub

<https://github.com/ies-research/multi-annotator-machine-learning>

