November 11, 2023

Leveraging Human-Machine Interactions for Computer Vision Dataset Quality Enhancement

Esla Timothy Anzaku, Hyesoo Hong, Jin-Woo Park, Wonjun Yang, Kangmin Kim, JongBum Won, Deshika Vinoshani Kumari Herath, Arnout Van Messem, Wesley De Neve

GHENT UNIVERSITY

FACULTY OF ENGINEERING AND ARCHITECTURE

GHENT UNIVERSITY GLOBAL CAMPUS

Center for Biosystems & Biotech Data Science

LIÈGE université Mathématique

Overview

The ImageNet Dataset

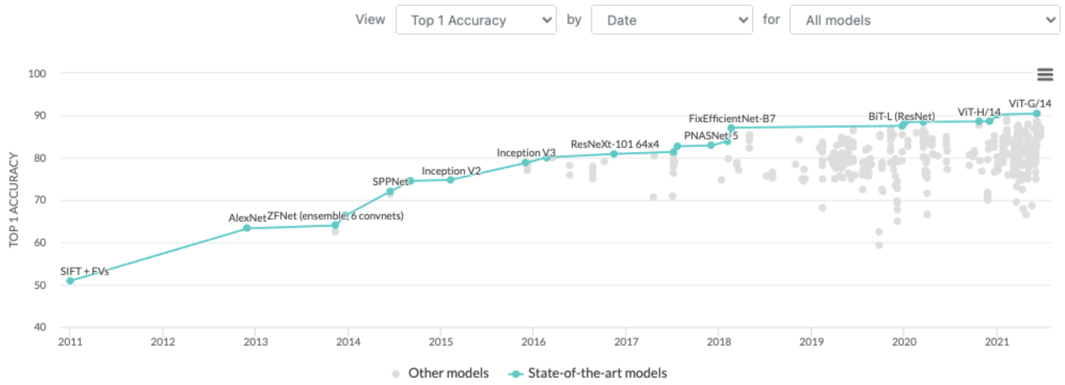
GHENT UNIVERSITY

Introducing ImageNet and ImageNet-1k

- **ImageNet**: Largest visual dataset for object recognition.
- Over 14 million images across approximately 22k categories.
- **ImageNet-**1k: A subset with 1k categories and over 1million images.
  - Used for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC).
  - Spans categories from ‘dogs’ and ‘plants’ to ‘building’ and ‘vehicles’
  - Central to major deep learning breakthroughs.
    - Example: Transfer Learning
  - Benchmark for model evaluation in computer vision.
    - Example: Supervised and Self-supervised Benchmarking

INTRODUCING IMAGENET AND IMAGENET-1*k*

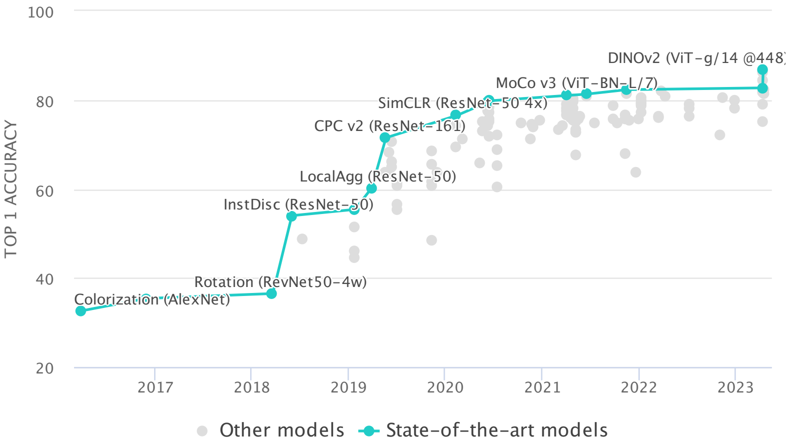
BENCHMARKING SUPERVISED IMAGE MULTI-CLASS CLASSIFICATION[1]



[1] Source: https://paperswithcode.com

INTRODUCING IMAGENET AND IMAGENET-1*k*

BENCHMARKING SELF-SUPERVISED IMAGE MULTI-CLASS CLASSIFICATION[2]



[2] Source: https://paperswithcode.com/sota/self-supervised-image-classification-on

The Single-Label Assumption

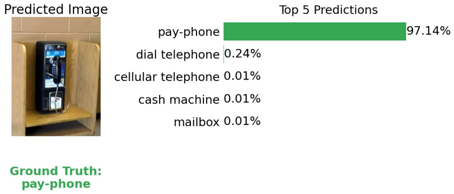
GHENT UNIVERSITY

The Single-label Assumption in ImageNet-1k

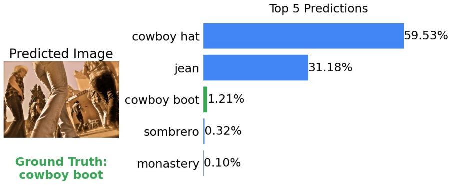
Implications for Metric Accuracy and Model Evaluation

- **Single-label Assumption**: Each image in ImageNet-1k is annotated with single label.
- Common metrics: *Top-1* and *Top-5* accuracies.
  - *Top-1 Accuracy*: The model's prediction matches the ground truth.
  - *Top-5 Accuracy*: The true label is among the model's top 5 predictions.
- **Assuming single-label correctness could skew evaluations, impacting not just top-1 and top-5 metrics but also Precision, Recall, ROC AUC, Negative Log Likelihood, ECE, and more.**

Top-1 correctness: Ground truth = topmost prediction



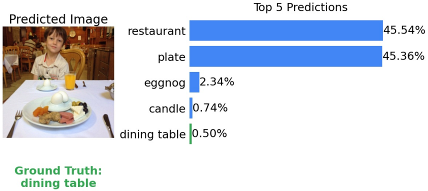
Top-5 correctness: Ground truth among topmost 5 predictions

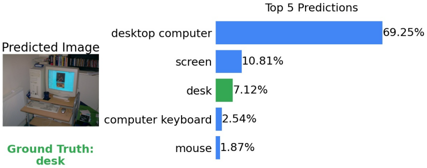


Why revisit the single-label assumption?

Why revisit the single-label assumption? (1/4)

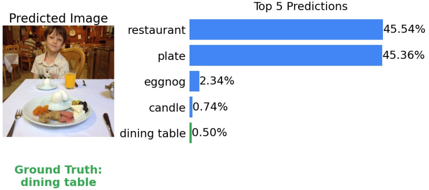
Qualitative Observations: Contrary Examples

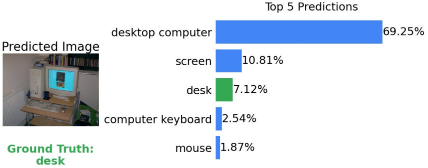


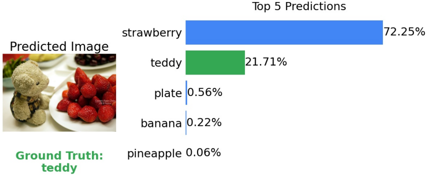


Why revisit the single-label assumption? (1/4)

Qualitative Observations: Contrary Examples

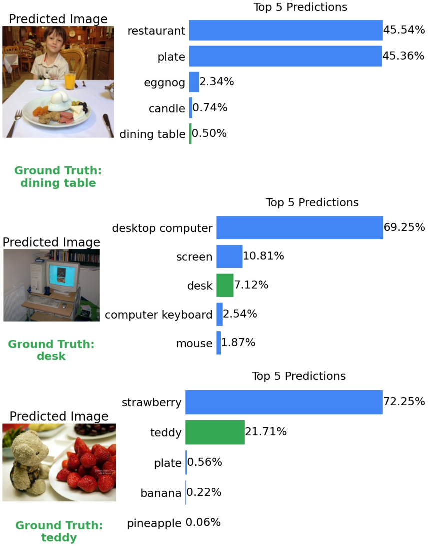






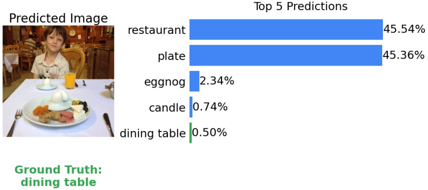
Why revisit the single-label assumption? (1/4)

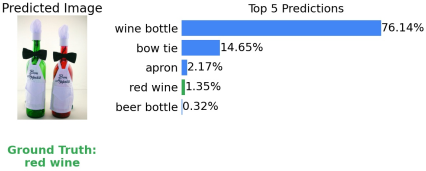
Qualitative Observations: Contrary Examples

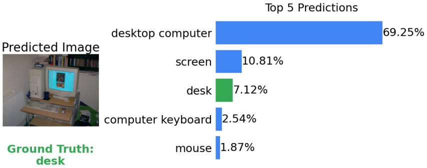


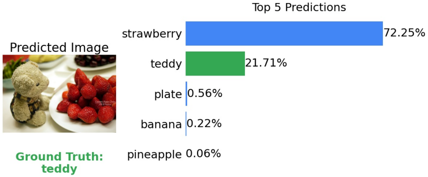
Why revisit the single-label assumption? (1/4)

Qualitative Observations: Contrary Examples



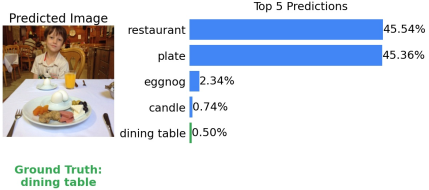


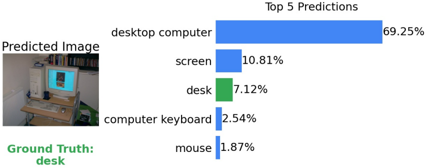


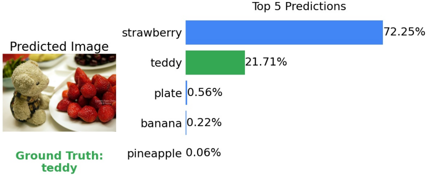


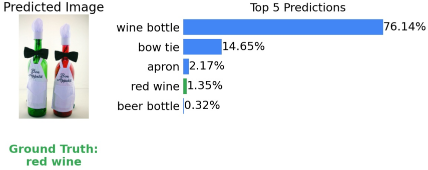
Why revisit the single-label assumption? (1/4)

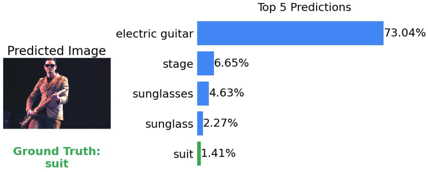
Qualitative Observations: Contrary Examples





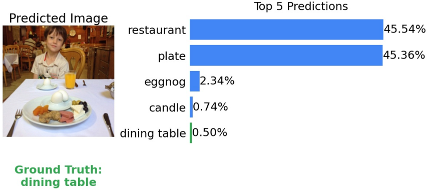


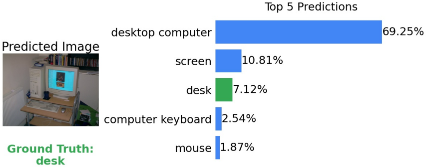


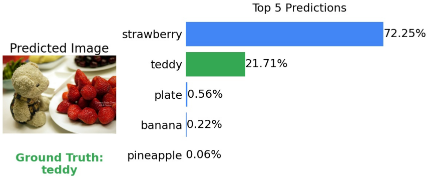


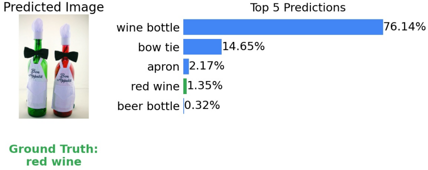
Why revisit the single-label assumption? (1/4)

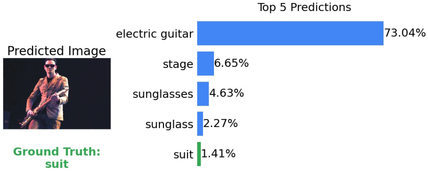
Qualitative Observations: Contrary Examples

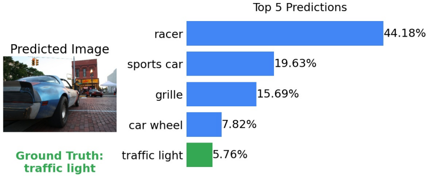






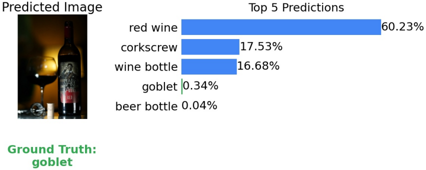






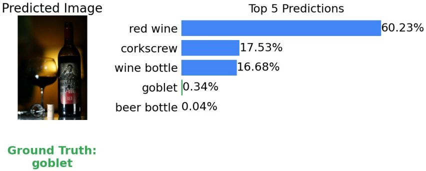
WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (1/4)

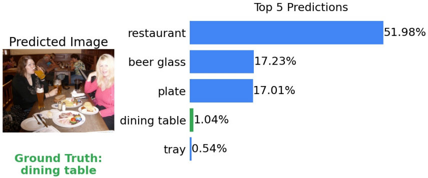
QUALITATIVE OBSERVATIONS: CONTRARY EXAMPLES



WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (1/4)

QUALITATIVE OBSERVATIONS: CONTRARY EXAMPLES

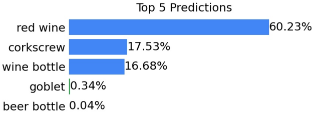




Why revisit the single-label assumption? (1/4)

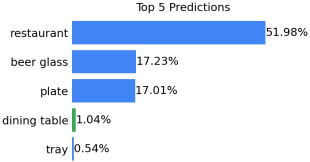
Qualitative Observations: Contrary Examples

Predicted Image



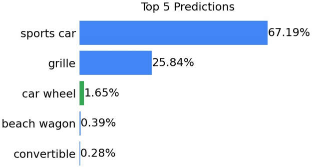
Ground Truth: goblet

Predicted Image



Ground Truth: dining table

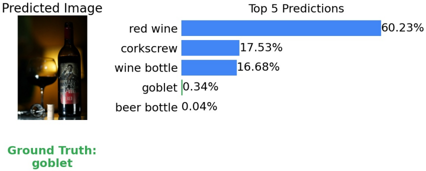
Predicted Image

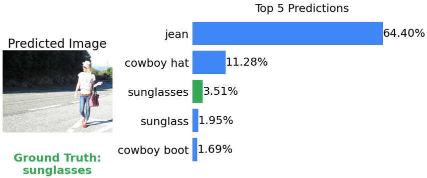


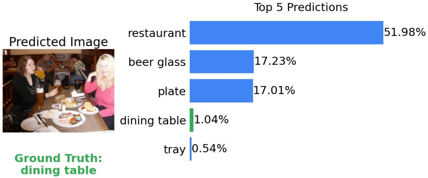
Ground Truth: car wheel

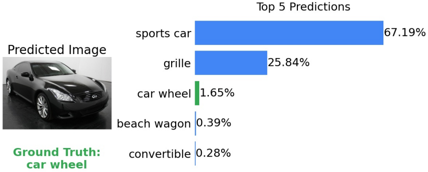
Why revisit the single-label assumption? (1/4)

Qualitative Observations: Contrary Examples



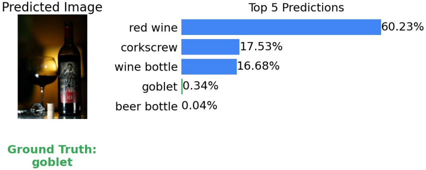


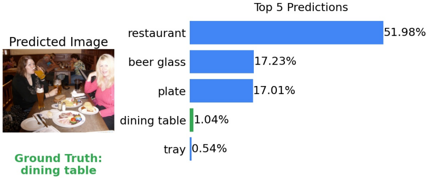


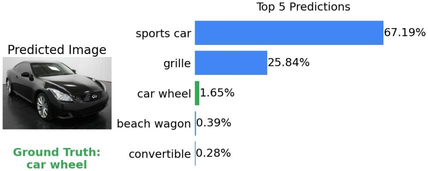


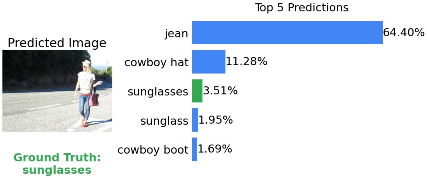
Why revisit the single-label assumption? (1/4)

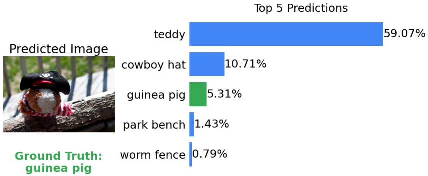
Qualitative Observations: Contrary Examples





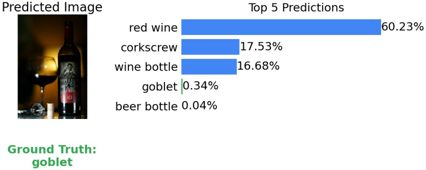


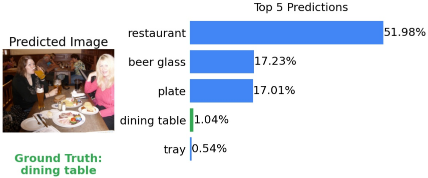


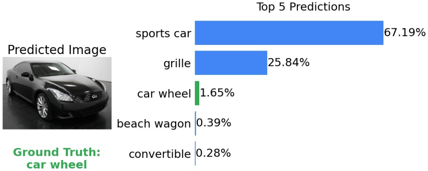


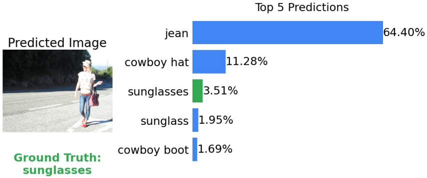
WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (1/4)

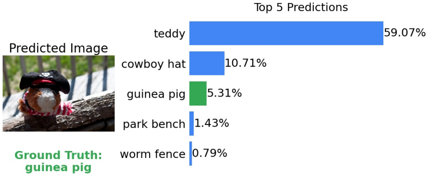
QUALITATIVE OBSERVATIONS: CONTRARY EXAMPLES

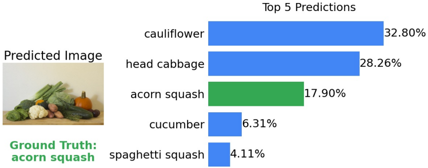






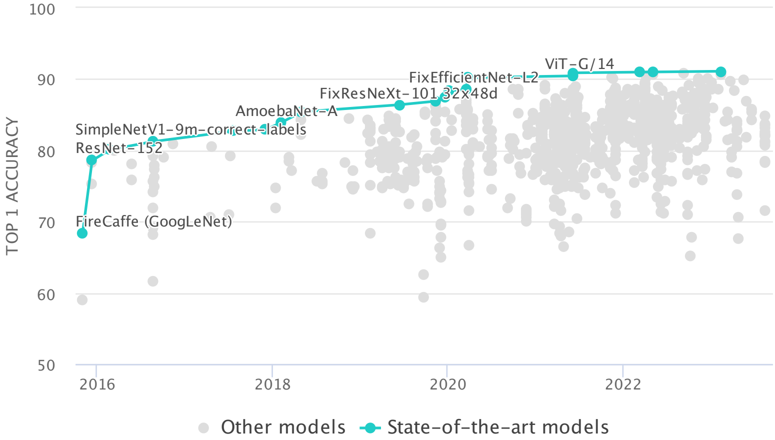






WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (2/4)

ACCURACY SATURATION: IS SOMETHING WRONG WITH THE DATA?[3]



Regardless of model architecture, training technique, dataset, and model size

[3] Source: https://paperswithcode.com/sota/image-classification-on-imagenet

WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (3/4)

UNEXPECTED ACCURACY DEGRADATION ON IMAGENET V2 DATASET

ImageNet validation set (50*k* Images)

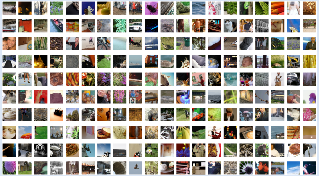
[a] Recht et. al., Do ImageNet Classifiers Generalize to ImageNet? (2019)

WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (3/4)

UNEXPECTED ACCURACY DEGRADATION ON IMAGENET V2 DATASET

ImageNet validation set (50k Images)

ImageNet V2[a] (10k Images)



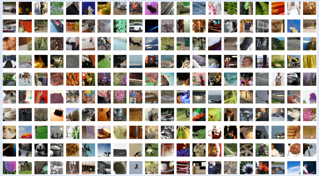
[a] Recht et. al., Do ImageNet Classifiers Generalize to ImageNet? (2019)

WHY REVISIT THE SINGLE-LABEL ASSUMPTION? (3/4)

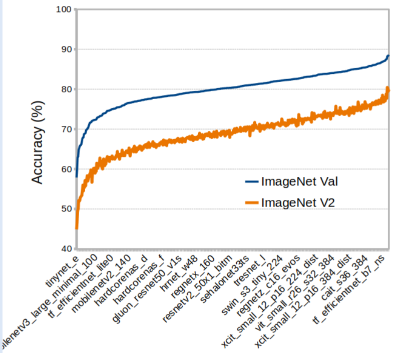
UNEXPECTED ACCURACY DEGRADATION ON IMAGENET V2 DATASET

ImageNet validation set (50*k* Images)

ImageNet V2[a] (10*k* Images)



Degradation Consistent Across 591 Models



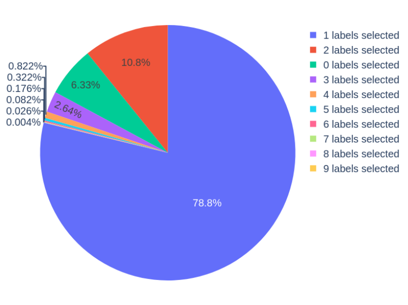
[a] Recht et. al., Do ImageNet Classifiers Generalize to ImageNet? (2019)

Why revisit the single-label assumption? (4/4)

Published work[4] on the Multi-label Nature of ImageNet Validation Set

- Reassessed ImageNet validation labels (50*k* images)
- **Task**: Identify all distinct objects in each image

Multi-label Proportions



[4] Source: Tsipras et. al., From ImageNet to ImageNet Classification: Contextualizing Progress on Benchmarks (2020).

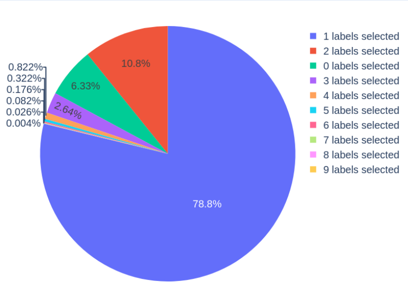
Why revisit the single-label assumption? (4/4)

Published work[4] on the Multi-label Nature of ImageNet Validation Set

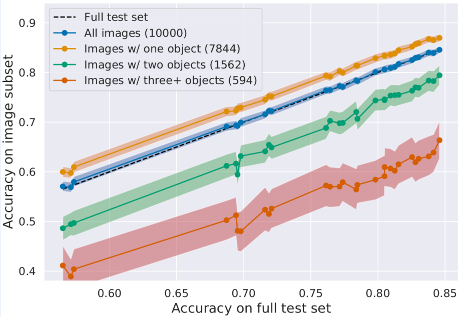
- Reassessed ImageNet validation labels (50*k* images)
- **Task**: Identify all distinct objects in each image

- Five annotators re-labeled the ImageNet-1*k* val. set
- **Full test set**: 50*k* images of the ImageNet validation set
- **All images**: 10k randomly selected images from the full val. set

Multi-label Proportions



Subset Accuracy



[4] Source: Tsipras et. al., From ImageNet to ImageNet Classification: Contextualizing Progress on Benchmarks (2020).

Our Framework – MultiLabelfy

GHENT
UNIVERSITY

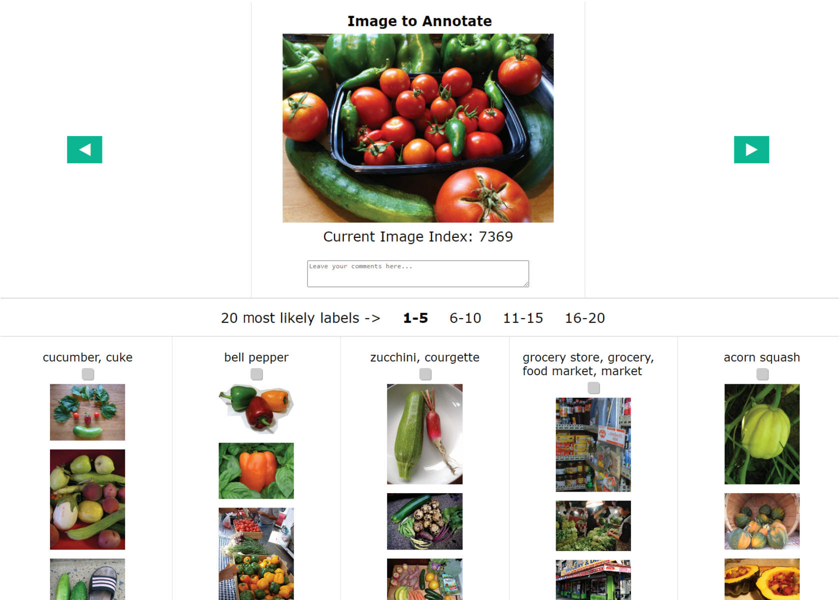
Why the need for MultiLabelfy?

Dataset Enhancement Challenges & Opportunities

- Annotation is labor-intensive and prone to errors
- Platforms like Mechanical Turk are often out of reach for smaller research groups
- A demand exists for open-sourced and rigorously reviewed dataset enhancement frameworks
- Available pre-trained models can be efficiently leveraged
- A user-friendly interface can greatly improve human-machine synergy

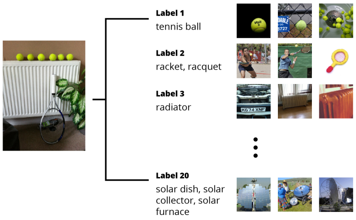
Our framework aims to leverage the opportunities while mitigating the challenges presented.

MultiLabelfy User Interface

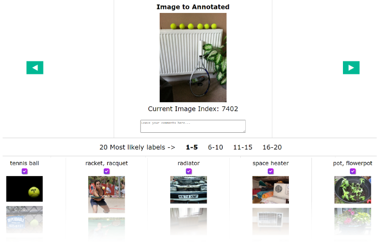


MULTILABELFY FRAMEWORK OVERVIEW

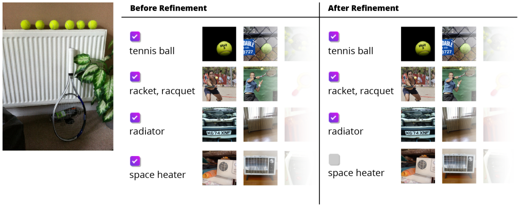
Label Proposal Generation



Human Multi-Label Annotation



Human Annotation Refinement



Annotation Disagreement Analysis

Annotator 1	Annotator 2	Disagreement
tennis ball racket, racquet	tennis ball racket, racquet	X
tennis ball racket, racquet space heater	tennis ball racket, racquet	O

Data Enhancement CaseStudy with Multilabelfy

Re-Labeling ImageNet V2: Experiment Summary

- **Stages1: Label Proposal Generation (Automated)**
  - Pre-trained Model Used: EVA-02[5] (**Top-1**: 90.05%; **Top-5**: 99.05%)
    - DNN Architecture: Vision Transformer
    - Trained on 38 million images
    - First fine-tuned on ImageNet-22*k* then fine-tuned on ImageNet-1*k*
  - Generates top 20 candidate labels per image

[5] Source: Sun et. al., A Visual Representation for Neon Genesis (2023)

Data Enhancement CaseStudy with Multilabelfy

Re-Labeling ImageNet V2: Experiment Summary

- **Stages1: Label Proposal Generation (Automated)**
  - Pre-trained Model Used: EVA-02[5] (**Top-1**: 90.05%; **Top-5**: 99.05%)
    - DNN Architecture: Vision Transformer
    - Trained on 38 million images
    - First fine-tuned on ImageNet-22*k* then fine-tuned on ImageNet-1*k*
  - Generates top 20 candidate labels per image
- **Stage 2: Human Multi-Label Annotation**
  - 14 annotators of various experience levels with computer vision and ImageNet dataset
  - All underwent training on the nuances of the task
  - Each image was annotated by two annotators

[5] Source: Sun et. al., A Visual Representation for Neon Genesis (2023)

Data Enhancement CaseStudy with MultilabelFy

Re-Labeling ImageNet V2: Experiment Summary

- **Stages1: Label Proposal Generation (Automated)**
  - Pre-trained Model Used: EVA-02[5] (**Top-1**: 90.05%; **Top-5**: 99.05%)
    - DNN Architecture: Vision Transformer
    - Trained on 38 million images
    - First fine-tuned on ImageNet-22*k* then fine-tuned on ImageNet-1*k*
  - Generates top 20 candidate labels per image
- **Stage 2: Human Multi-Label Annotation**
  - 14 annotators of various experience levels with computer vision and ImageNet dataset
  - All underwent training on the nuances of the task
  - Each image was annotated by two annotators
- **Stage3: Annotation Disagreement Analysis (Automated)**
  - 6, 425 images were selected for the next refinement stage

[5] Source: Sun et. al., A Visual Representation for Neon Genesis (2023)

Data Enhancement CaseStudy with Multilabelfy

Re-Labeling ImageNet V2: Experiment Summary

- **Stages1: Label Proposal Generation (Automated)**
  - Pre-trained Model Used: EVA-02[5] (**Top-1**: 90.05%; **Top-5**: 99.05%)
    - DNN Architecture: Vision Transformer
    - Trained on 38 million images
    - First fine-tuned on ImageNet-22*k* then fine-tuned on ImageNet-1*k*
  - Generates top 20 candidate labels per image
- **Stage 2: Human Multi-Label Annotation**
  - 14 annotators of various experience levels with computer vision and ImageNet dataset
  - All underwent training on the nuances of the task
  - Each image was annotated by two annotators
- **Stage3: Annotation Disagreement Analysis (Automated)**
  - $6,425$ images were selected for the next refinement stage
- **Stage 4: Human Annotation Refinement**
  - 5 annotators participated; 4 of them had participated in Stage 2.
  - Only 129 of 10*k* images remained unlabeled after this stage.

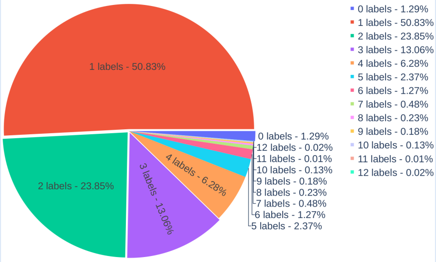
[5] Source: Sun et. al., A Visual Representation for Neon Genesis (2023)

Data Enhancement CaseStudy with Multilabelfy

Re-Labeling ImageNet V2: Key Results

- **About 50% images have more than one valid label**

Multi-label Proportions

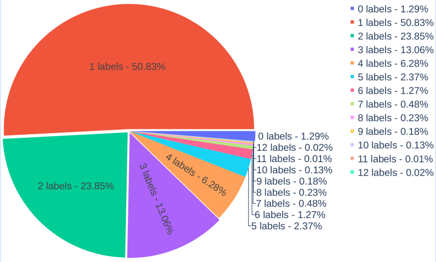


Data Enhancement CaseStudy with MultiLabelfy

Re-Labeling ImageNet V2: Key Results

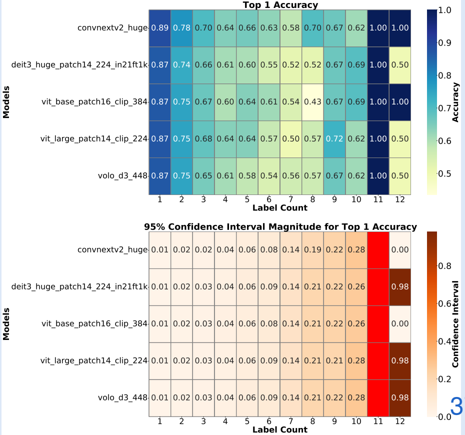
- **About 50% images have more than one valid label**

Multi-label Proportions



- **Label count negatively correlates with top-1 accuracy**

Top-1 Accuracy versus Label Count



Final Thoughts

GHENT UNIVERSITY

Reevaluating the Single-label Assumption

Why Embracing Multi-label Realities Matters

- **To Reflect Real-World Complexities**
  - Ensure future labeling reflects real-world complexities
  - Our DNN models are already hinting at the disconnect

Reevaluating the Single-label Assumption

Why Embracing Multi-label Realities Matters

- **To Reflect Real-World Complexities**
  - Ensure future labeling reflects real-world complexities
  - Our DNN models are already hinting at the disconnect
- **To Enhance Model Evaluation**
  - Capture true model capabilities without bias
  - Prevent unfair penalization of models for valid alternative predictions

Reevaluating the Single-label Assumption

Why Embracing Multi-label Realities Matters

- **To Reflect Real-World Complexities**
  - Ensure future labeling reflects real-world complexities
  - Our DNN models are already hinting at the disconnect
- **To Enhance Model Evaluation**
  - Capture true model capabilities without bias
  - Prevent unfair penalization of models for valid alternative predictions
- **To Inform Data Collection and Labeling**
  - Advocate for datasets that allow DNNs to demonstrate their full potential
  - Encourage the incorporation of a broader spectrum of labels

Reevaluating the Single-label Assumption

Why Embracing Multi-label Realities Matters

- **To Reflect Real-World Complexities**
  - Ensure future labeling reflects real-world complexities
  - Our DNN models are already hinting at the disconnect
- **To Enhance Model Evaluation**
  - Capture true model capabilities without bias
  - Prevent unfair penalization of models for valid alternative predictions
- **To Inform Data Collection and Labeling**
  - Advocate for datasets that allow DNNs to demonstrate their full potential
  - Encourage the incorporation of a broader spectrum of labels
- **To Fuel Progress in the Field**
  - Foster innovation with more accurate and holistic model evaluations

Reevaluating the Single-label Assumption

Why Embracing Multi-label Realities Matters

- **To Reflect Real-World Complexities**
  - Ensure future labeling reflects real-world complexities
  - Our DNN models are already hinting at the disconnect
- **To Enhance Model Evaluation**
  - Capture true model capabilities without bias
  - Prevent unfair penalization of models for valid alternative predictions
- **To Inform Data Collection and Labeling**
  - Advocate for datasets that allow DNNs to demonstrate their full potential
  - Encourage the incorporation of a broader spectrum of labels
- **To Fuel Progress in the Field**
  - Foster innovation with more accurate and holistic model evaluations
- **To Boost Reliability and Trust**
  - Promote rigorous validation for consistent real-world performance
  - Establish more reliable benchmarks to inspire stakeholder confidence

Future Research Interests

- What are the costs of the single-label assumption?
  - How does this assumption contribute to the **surprising brittleness** of DNN models?
  - What are the costs of utilizing **powerful models** on **simplified assumptions**?
  - To what extent does the single-label assumption foster **overfitting** to **dataset idiosyncrasies**?
  - Could challenging the single-label assumption stimulate a renewed discussion on **nuanced model evaluation**?

Esla Timothy Anzaku
eslatimothy.anzaku@ugent.be