Goal: Build a toolkit for reliable and robust real-world ML deployment

Revisiting the underpinnings of the ML framework

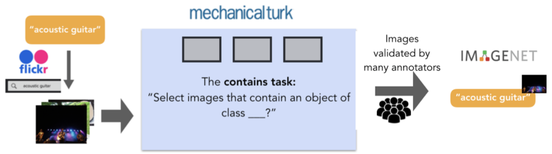
To what extent does our **conceptual view** of the ML framework explain how **models behave in practice**?

I. Algorithms and architectures [SSM'18; STM'18; IESTJRM'20]
II. Datasets and benchmarks [TSEIM'20; EISTM'20]

Benchmark-Task Misalignment

→ The need for *scalability* changes how we collect data

→ The resulting data collection practices introduce biases in annotations, causing them to *deviate from the ground truth*



→ "Gold-standard treatment" of datasets is *problematic*: leads to spurious associations in models and a skewed view of performance

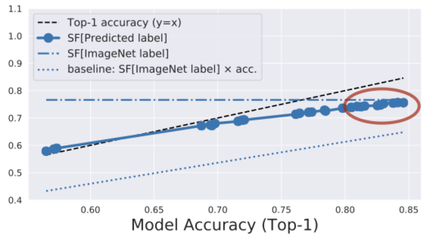
ImageNet label: *"banana"*

Our annotations

Main object label: *"wine bottle"*

Other objects:
- *"banana"*
- *"lemon"*
- *"orange"*
- *"corkscrew"*
- *"red wine"*

ImageNet: "pickelhaube"
Annotators: "military uniform"



Annotators can't differentiate ImageNet labels from model predictions
Have we reached the limit of ImageNet's usefulness?

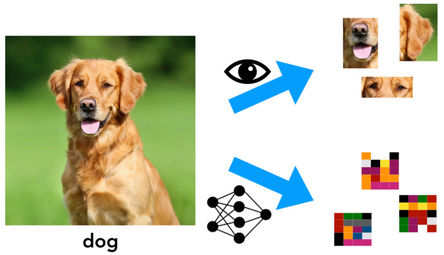
Diagnosing and correcting model failures

Can models perform well **beyond curated environments**? What happens if they fail to do so?

I. Adversarial vulnerability [SSTTM'18; ISTETM'19; STTIEM'19; TSM'19; EISTTM'19]]
II. Natural distribution shifts [STM'20]

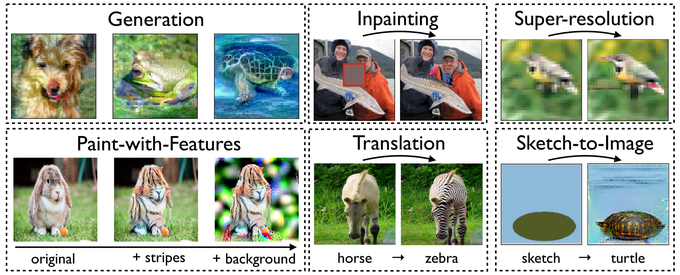
How do models make decisions?

→ *Human-ML misalignment*: models succeed by using inherently different features than the ones used by humans



→ Adversarial examples are a direct manifestation of this

→ *Robust models* learn features that are more perceptually-aligned, and are also *better suited for downstream tasks*



Adversarial robustness goes beyond security:
Could it serve as a prior for building better models?

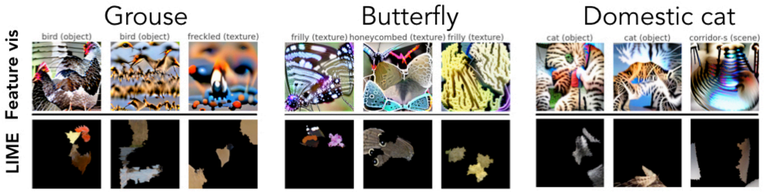
Towards a holistic view of the ML pipeline

Rethinking dataset design: How can we build realistic *and representative* datasets in a *scalable* manner?

→ Identify biases in data collection via *human-in-the-loop studies*

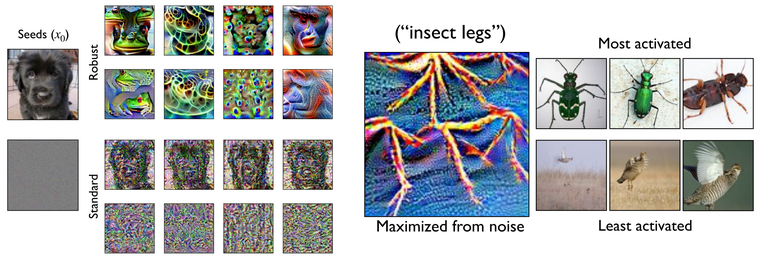
→ Develop training/evaluation procedures that can alleviate dataset deficiencies and artifacts

Feature debugging: How do we *detect undesirable dependencies* learned by models?



Identifying a sparse set of causal features in standard models [WSM'20]

Feature engineering: Can we use the *robustness* framework to *control spurious correlations* in models?



Read More: gradientscience.org