Task Discovery: Finding the Tasks that Neural Networks Generalize on

Andrei Atanov, Andrei Filatov, Teresa Yeo, Ajay Sohmshetty, Amir Zamir

taskdiscovery.epfl.ch

EPFL

Task Discovery Problem

- Neural Networks can fit any task, i.e., any labeling of a set of images [1,2], but what tasks can they generalize on?

Task Discovery: finds **different generalizable** tasks **automatically**.

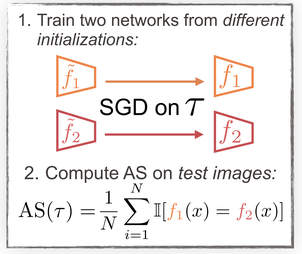
- How do these tasks *look*? What do they *indicate*?
- Generalizable tasks reflect the inductive biases of NNs
  ⇒ can help us understand deep learning better.

Generalizability via Agreement Score

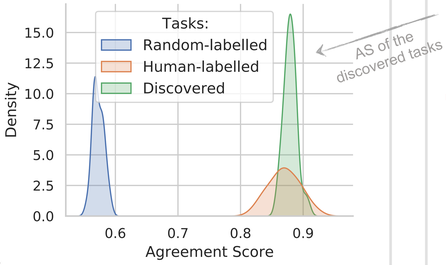
- How do we differentiate between generalizable (e.g., human-labeled) and non-generalizable (e.g., random-labeled) tasks quantifiably?

A Human-Labeled Task | A Random-Labeled Task

- We use **Agreement Score (AS)**, a generalization-based quantity, to identify generalizable tasks computationally:



Agreement Score (AS) Computation

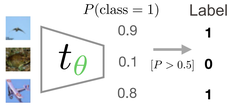


Automatic Discovery via Meta-Optimization

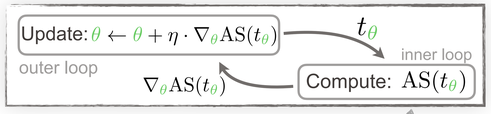
- We optimize the labeling of a set of images, resulting in a high-AS discovered task:

$$\tau_{\mathrm{d}} \leftarrow \arg\max_{\tau} \ \mathrm{AS}(\tau)$$

- Model the labeling with a task network t_θ that outputs the label for each image.



- Use meta-optimization:
  - **Outer loop:** update the task network that provides labels
  - **Inner loop:** compute the AS of a given task



Agreement Score Meta-Optimization

Find Different Tasks via Dissimilarity Constraint

- Meta-optimization results in a single task, but there could be many high-AS tasks we want to discover.
- We aim to discover a set of dissimilar tasks:

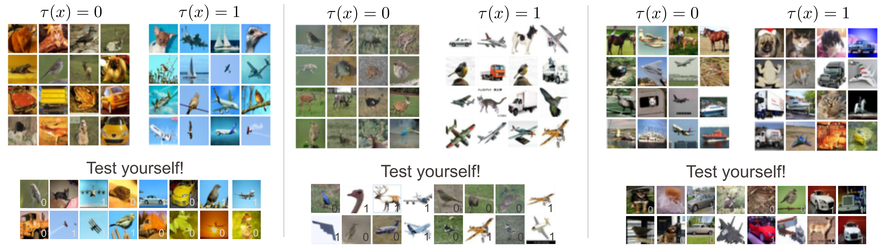
$$\underset{T=\{t_{\theta_1},\ldots,t_{\theta_K}\}}{\arg\max} \quad \mathbb{E}_{t_\theta \sim T} \mathrm{AS}(t_\theta) - \lambda \cdot L_{\mathrm{sim}}(T)$$

The set of discovered high-AS tasks

Dissimilarity loss to avoid identical tasks

- Use a shared encoder with a uniformity loss in practice:

Visualisation of the Discovered Tasks

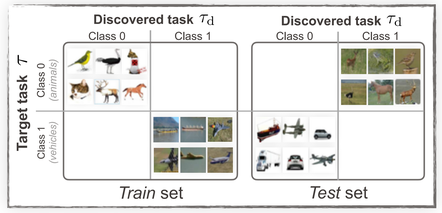


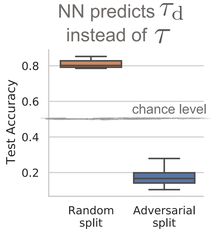
- Some tasks are *not easy to interpret visually* by humans, and **are not meant to be**!
  (see Sec. 5.4 of the paper)

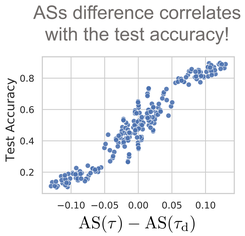
Discovered Tasks Reveal NNs failure modes via Adversarial Splits

- Discovered tasks reflect inductive biases of NNs and statistics patterns in data.
- Adversarial splits use them to highlight NNs' failure modes by creating a spurious correlation between a discovered and target to "fool" the network:

Adversarial** train-test **split







Summary

- Find generalizable tasks automatically via AS meta-optimization.
- Discovered tasks reflect NNs' inductive biases and statistical patterns in data ⇒ can help us analyze and understand them better.
  - Example: adversarial splits reveal NNs' failure modes.

References:
[1] Understanding deep learning requires rethinking generalization, *Zhang et al., ICLR 2017*
[2] Small ReLU networks are powerful memorizers: a tight analysis of memorization capacity, *Yun et al., NeurIPS 2019*