How much human-like visual experience do current SSL algorithms need to achieve human-level object recognition?

Emin Orhan (eo41@nyu.edu)

In a nutshell

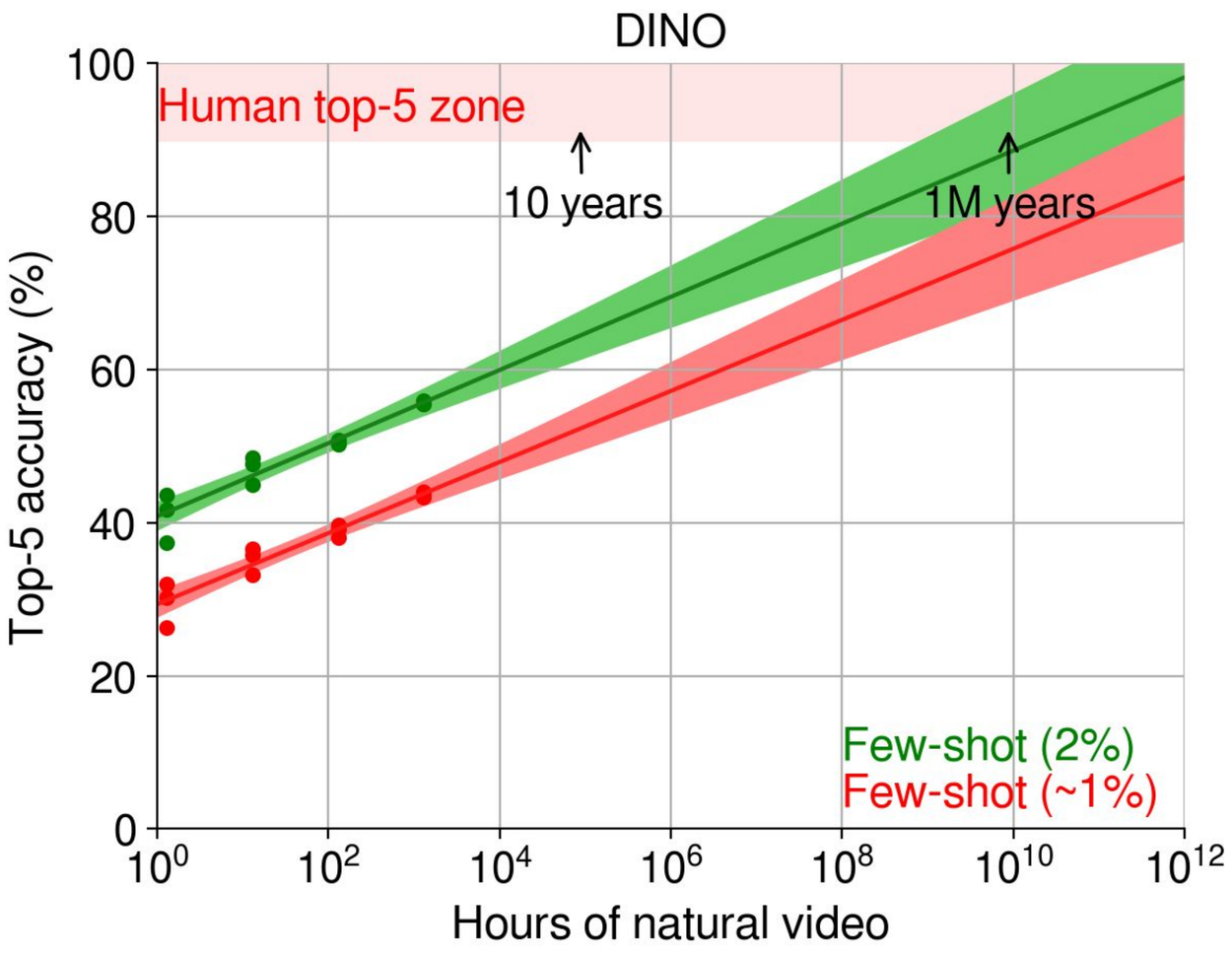
Q: Can our current SSL algorithms achieve human-level object recognition if given the same kind and amount of visual experience as humans?

A: No. They need orders of magnitude more "human-like" data than humans.

Human-like visual experience

- Combined 5 video datasets for a total of **1301 hours** (~54 days) of human-like natural video data
- Videos are continuous, temporally extended (each ~0.5-1 hour long), mostly egocentric (headcam)

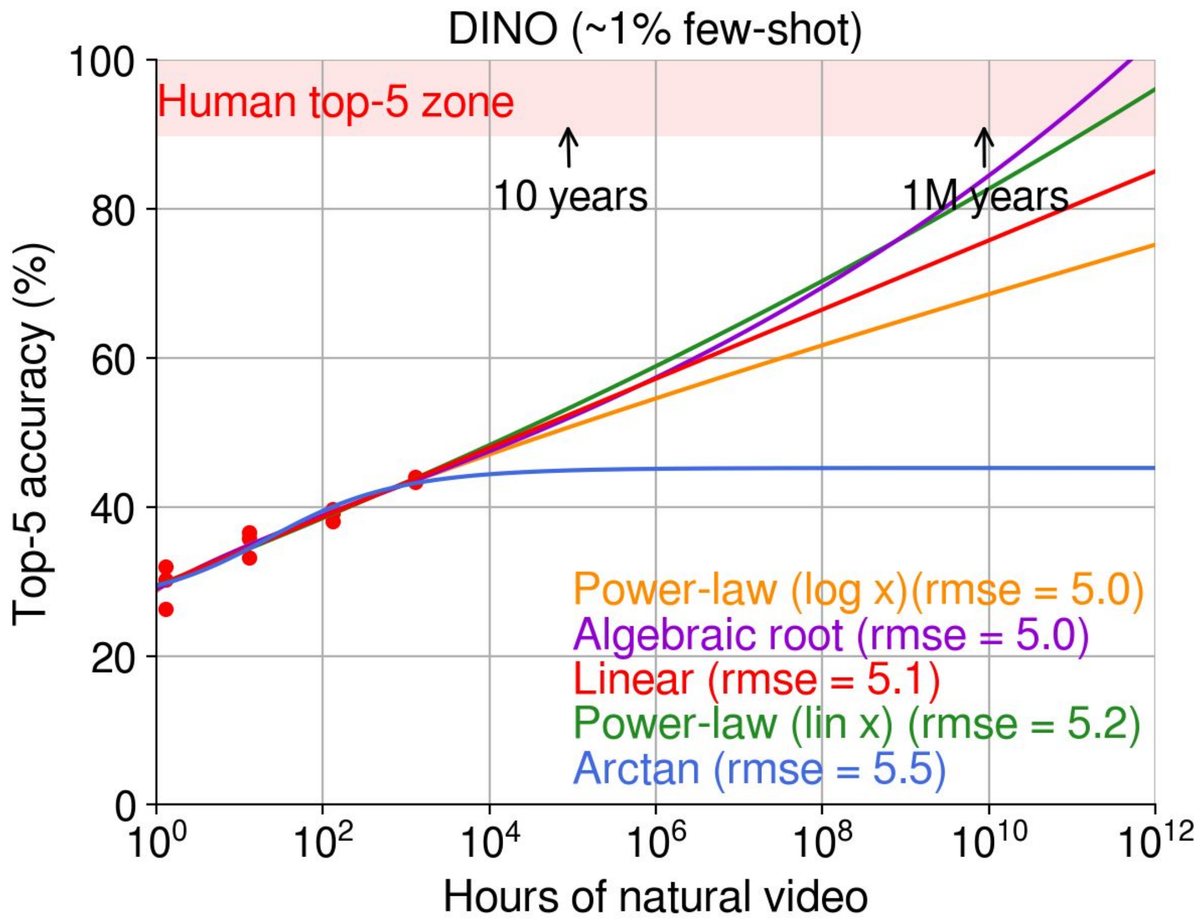
Achieving human-level accuracy on ImageNet



Amount of natural video needed to reach human-level accuracy (in years):

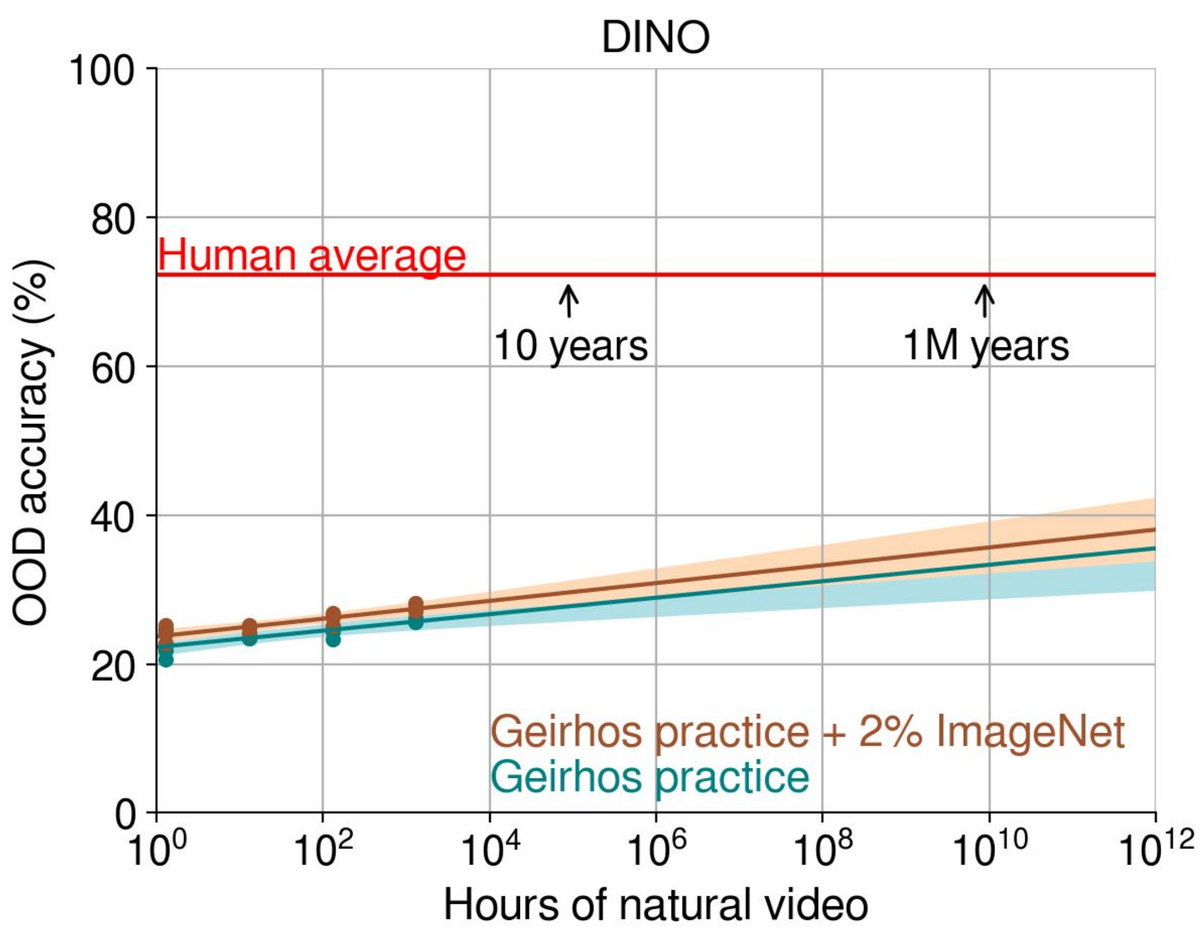
eval. method	DINO
few-shot (~1%)	1.4G (30.0M, 0.3T)
few-shot (2%)	2.3M (0.1M, 0.2G)

Alternative extrapolation functions



Achieving human-level robustness on ImageNet

Geirhos et al., *NeurIPS 2021*



Interpretation

Three possibilities:

- Minor variations on current algorithms will suffice (empiricist)
- Embodiment, multimodality needed (empiricist)
- More substantive inductive biases needed (rationalist)