Comparing supervised learning dynamics: Deep neural networks match human data efficiency but show a generalisation lag

Lukas S. Huber[1,2], Fred W. Mast[2], Felix A. Wichmann[1]

Most human-to-DNN behavioral comparison studies of object recognition focus on final outcomes, not on the learning process.

Empirical side-by-side comparison of supervised representation learning in humans and DNNs with aligned learning conditions.

Results indicate that while DNNs match humans in data efficiency, they show a pronounced generalization lag, unlike humans.

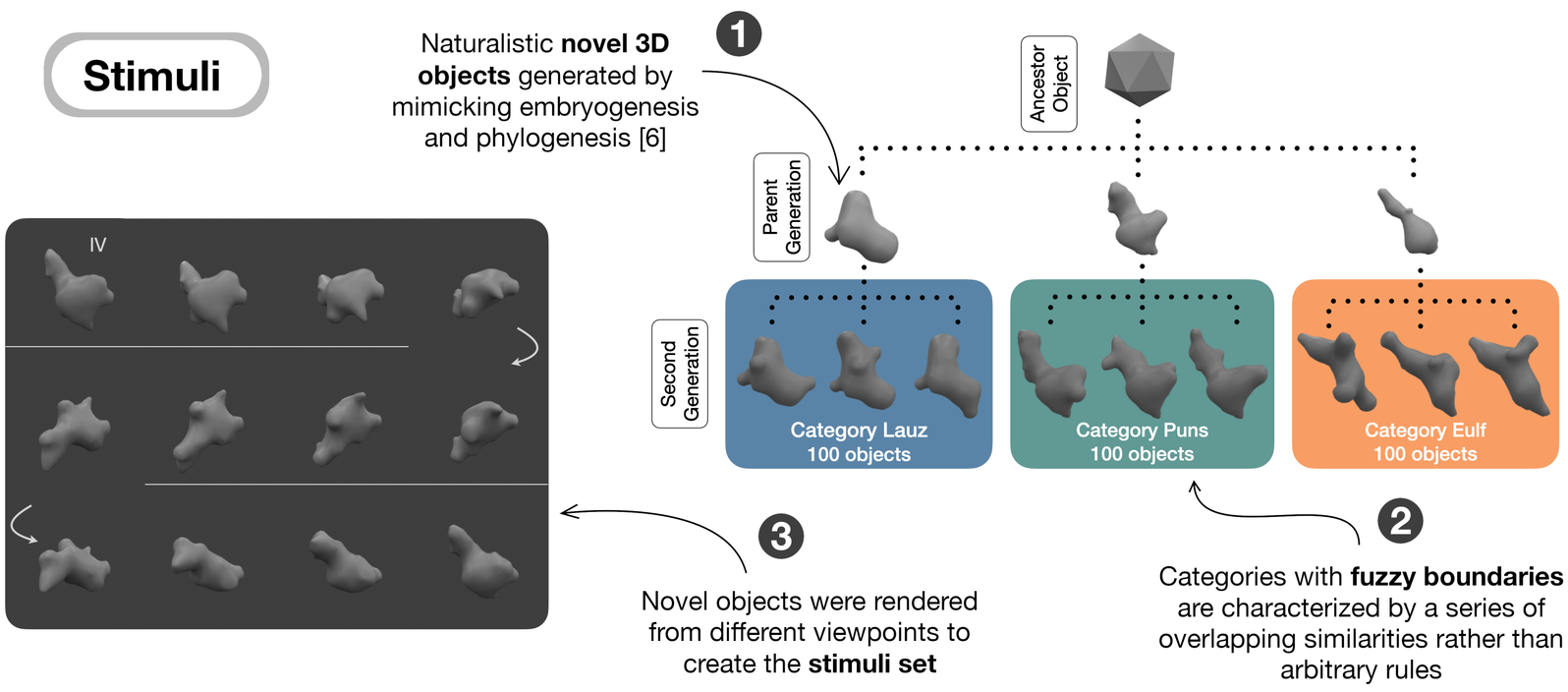
Motivation

Despite deep neural networks (DNNs) surpassing human-level performance in object recognition tasks [1], more fine-grained behavioral assessments reveal significant representational divergences, as evidenced by differences in robustness [2], error patterns at the category and trial level [3], susceptibility to adversarial attacks [4], and crowding [5]. Identifying the origins of these representational differences is a non-trivial task, given the vast search space for potential causes.

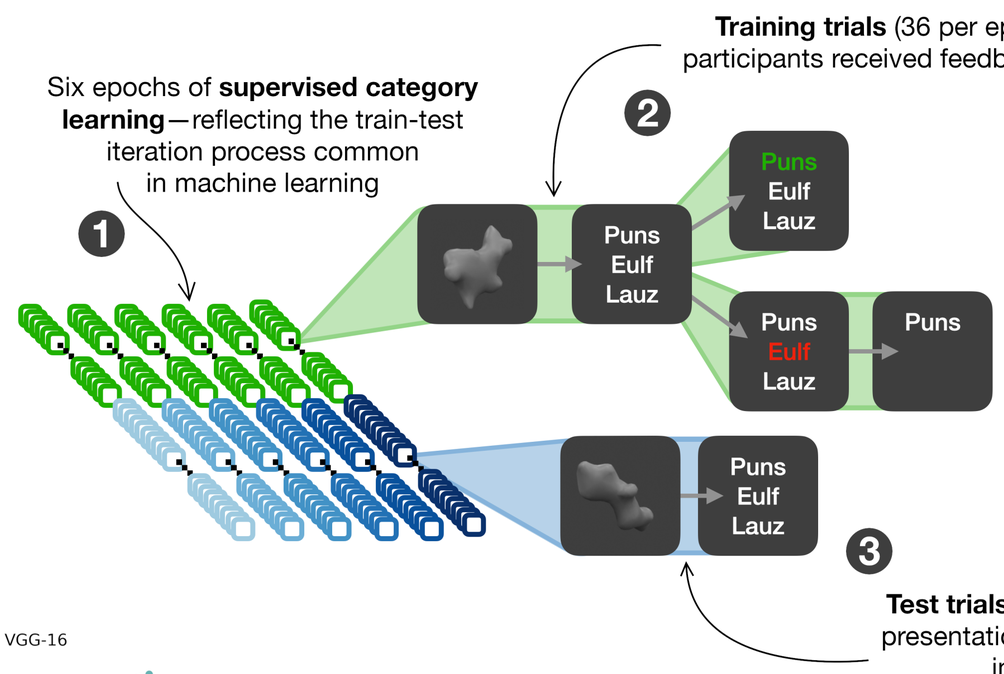
Rationale

Comparing the process in which representations are acquired appears crucial: for each learning condition aligned during representation acquisition, the search space becomes more constrained. We investigate how humans and DNNs acquire novel representations in a constrained environment with aligned learning conditions, tracking behavioral changes across six epochs of a category learning task. At each epoch, we assess the generalization of learned representations to previously unseen data.

Stimuli



Learning environment

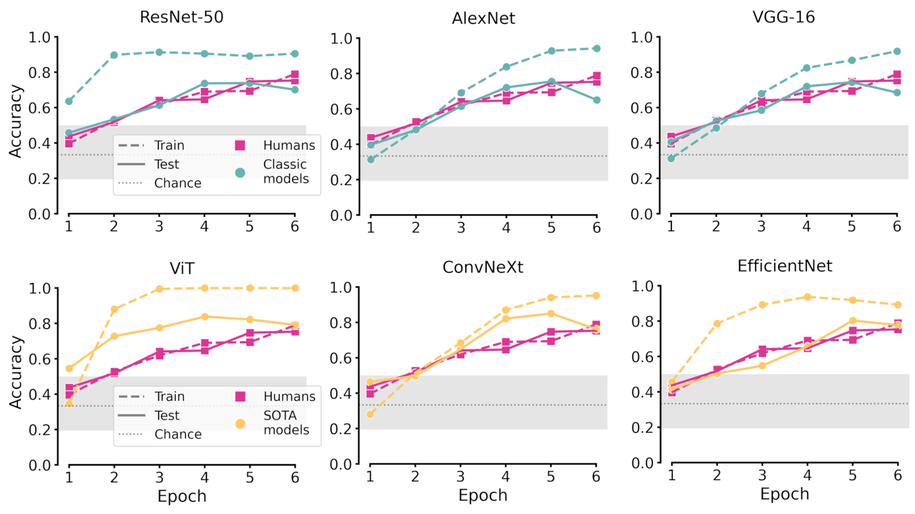


Aligned learning conditions

The **learning goal** is for both humans and DNNs to acquire representations of novel 3D objects. To better mimic human **prior knowledge** and ensure a fair starting point, DNNs are pre-trained on ImageNet. Consistent **input data** is assured by providing both DNNs and humans with the **exact same training and test data** in a matched supervised learning environment.

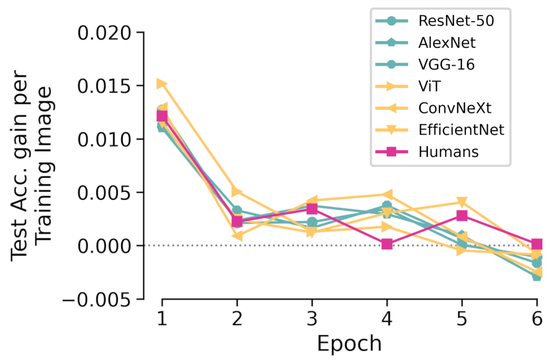
Results

1



2

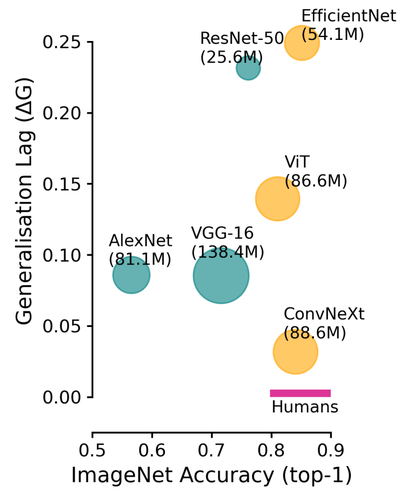
DNNs are not inherently less data efficient compared to humans. As quantification of data efficiency, we calculate the average test accuracy gain per training image at each epoch. Across all epochs, humans, CNNs and SOTA models show comparable data efficiency. Data efficiency is initially high during the early stages of learning but gradually approaches zero as learning progresses.



3

While SOTA models outperform classic CNNs in terms of ImageNet accuracy, this does not imply more human- aligned generalisation behavior. Generalisation lag is quantified as the mean difference between training and test accuracy, selectively focusing on epochs where learning occurred—excluding epochs where training performance did not significantly exceed chance level, and epochs where models showed signs of overfitting.

$$\Delta G = \frac{1}{|E|} \sum_{i \in E} (acc_{train,i} - acc_{test,i})$$



Paper

Poster

Increasing test performance indicates that both, humans and DNNs, learn novel generalizable representations. Notably, DNNs match or exceed human test performance indicating comparable data efficiency**.** However, while humans immediately form generalizable representations, most DNNs show a pronounced generalisation lag. Each model's performance is averaged across 20 fine-tuning runs and human trajectories are averaged over observers.

Cotntact
lukas.s.huber@unibe.ch
@LukasShuber

[1] He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision*.

[2] Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems*, 31.

[3] Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33.

[4] Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, 10(1), 1334.

[5] Lonnqvist, B., Bornet, A., Doerig, A., & Herzog, M. H. (2021). A comparative biology approach to DNN modeling of vision: A focus on differences, not similarities. *Journal of Vision*, 21(10), 17-17.

[6] Hauffen, K., Bart, E., Brady, M., Kersten, D., & Hegdé, J. (2012). Creating objects and object categories for studying perception and perceptual learning. *JoVE (Journal of Visualized Experiments)*, (69), e3358.

1 UNIVERSITÄT BERN
2 EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Funding: Swiss National Science Foundation

ICLR 2024 Workshop on Representational Alignment

ICLR