Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualizations

Judy Borowski*, Roland S. Zimmermann*, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis‡, Matthias Bethge‡, Wieland Brendel‡

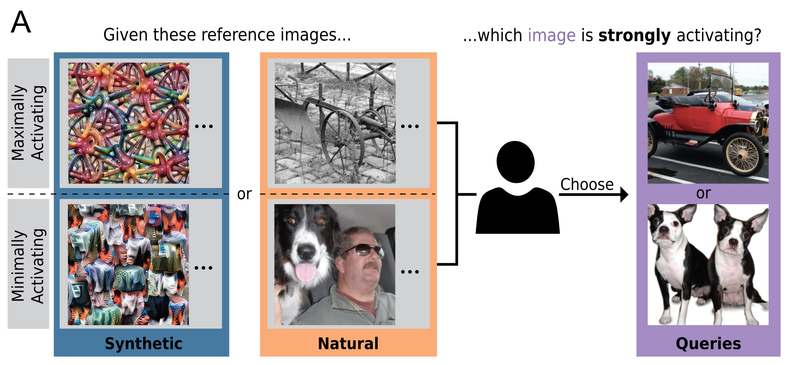
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

TL;DR: Using human psychophysical experiments, we show that natural images can be significantly more informative for interpreting neural network activations than a synthetic feature visualization.

Motivation

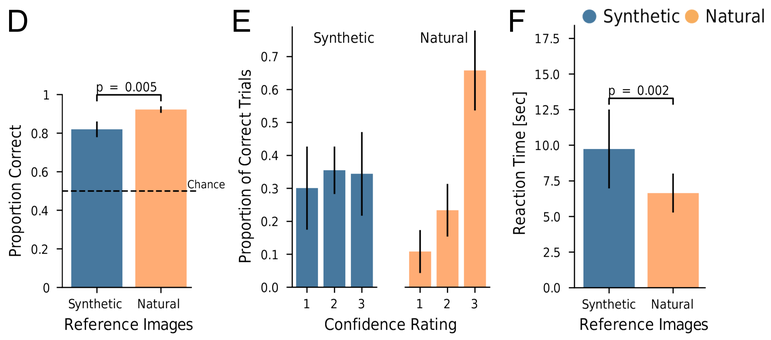
Feature visualizations such as synthetic maximally activating images are a popular explanation method. They are used to better understand the information processing of convolutional neural networks (CNNs). At the same time, there are concerns that these visualizations might not accurately represent CNNs' inner workings. Here, we investigate how much extremely activating images help humans to predict CNN activations.

Human experiments

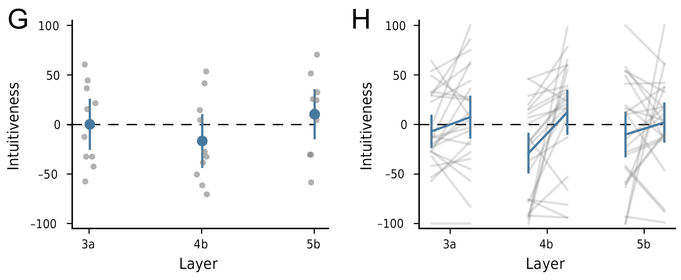


- Well-controlled lab experiments
- 2 experiments: 33 participants
- Feature visualizations by Olah et al.[1]
- Baseline: natural images (ImageNet)
- Network: InceptionV1[2]

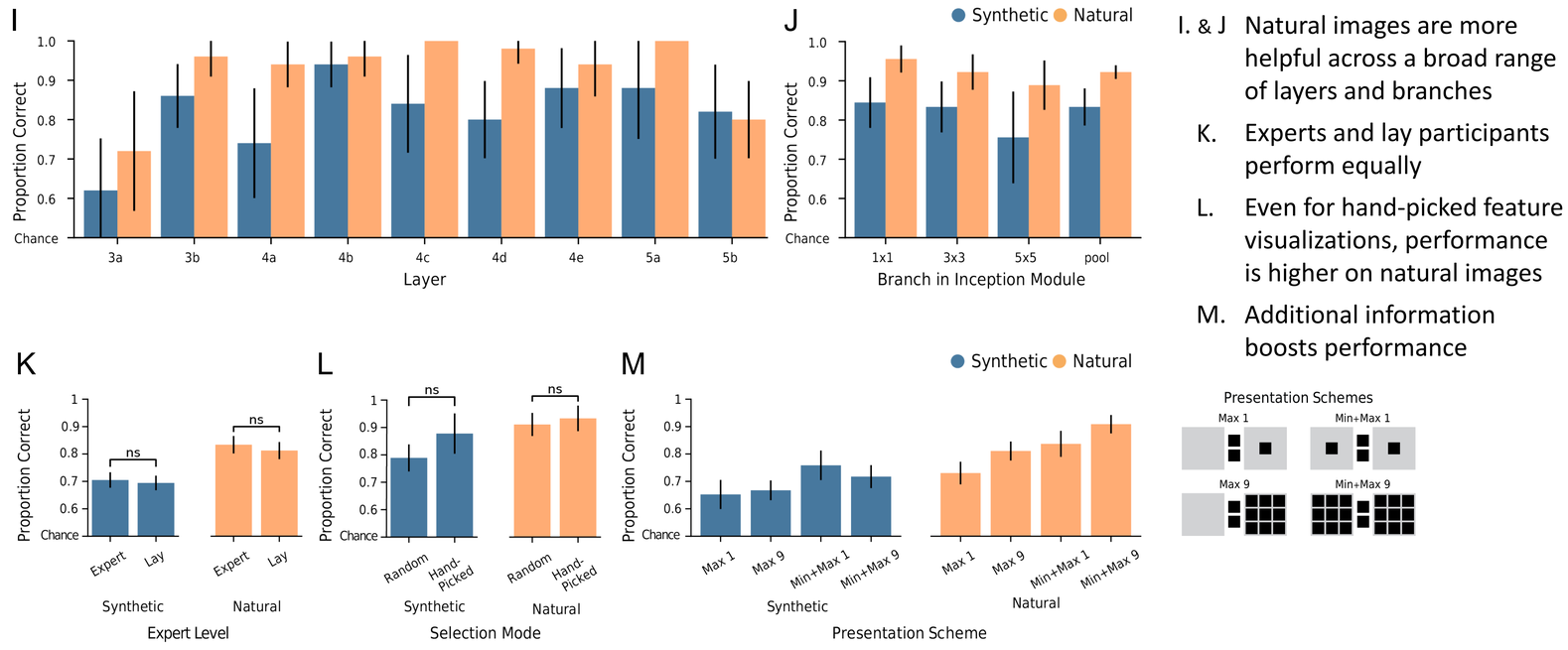
Better, more confident and faster with natural images



Subjectively, interpretability varies greatly



The superiority of natural images (mostly) holds across layers, branches and various conditions



I. & J. Natural images are more helpful across a broad range of layers and branches

K. Experts and lay participants perform equally

L. Even for hand-picked feature visualizations, performance is higher on natural images

M. Additional information boosts performance

Conclusion

- Synthetic images provide humans with helpful information about feature map activations
- However, exemplary natural images are even more helpful
- Mostly holds across the network and various conditions
- Our results highlight the need for human quantitative evaluations of feature visualizations

References, Funding and Code

[1] Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. "Feature visualization." Distill 2.11 (2017): e7.
[2] Szegedy, Christian, et al. "Going deeper with convolutions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

Bundesministerium für Bildung und Forschung
imprs-is
ROBUST VISION
DFG

bit.ly/2P36T72