Feature Visualizations do not sufficiently explain hidden units of Artificial Neural Networks

Thomas Klein, Wieland Brendel and Felix A. Wichmann

Feature Visualizations are supposedly useful for understanding ANNs.

We show that they are less interpretable than natural images.

This effect gets stronger for later layers of the network.

Motivation

ANNs are highly predictive tools widely employed in vision science. But good models should not only be predictive, but also aid understanding. [1]

❓ What do units in ANNs learn?

💡 Look at highly activating images!

... Indigo Finch? birds? bird perched on stick? green background? something else?

💡 Synthetically generate maximally activating images! [2]

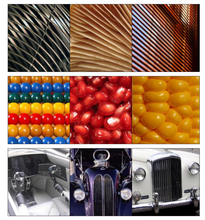
... Can you recognize what this is supposed to be?

Maybe, units are **polysemantic**, i.e. sensitive to multiple unrelated concepts or patterns that get "mixed" in the synthetic images.

Methods

Randomly select units from early, middle and late layers of GoogLeNet.
Find top 10 most activating natural images.
Generate 10 max. activating Feature Visualizations.

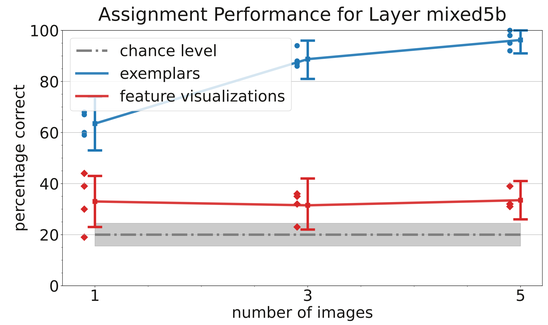
Exemplar Trial:
- Select 3 units
- Show a row of images for each unit
- Split them in half and shuffle
- Participants have to match correctly

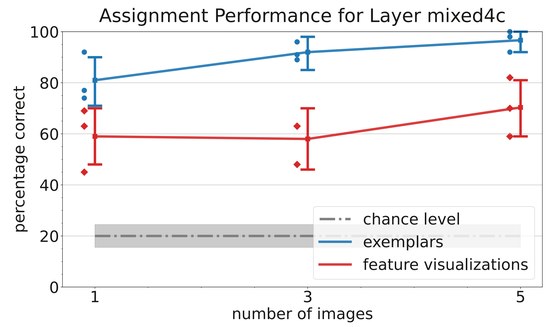


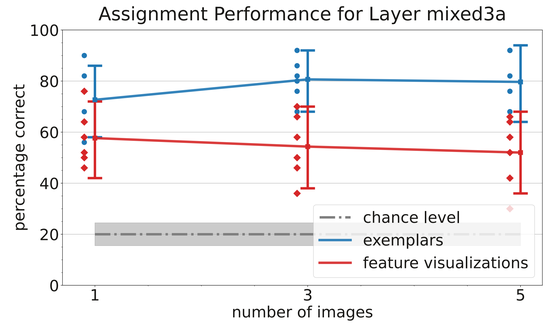
Feature Visualization Trial:
- Select 3 units
- Show 3 natural images for each unit
- Show 3 Feature Visualizations
- Participants have to match correctly

Experimental Conditions:
- Number of images = {1, 3, 5}
- Layer in the network = {early, middle, late}

Results







References
[1] Wichmann & Geirhos. Annual Review of Vision Science, 2023.
[2] Olah, Mordvintsev and Schubert. Distill, 2017.

Acknowledgements
This work has been supported by the Machine Learning Cluster of Excellence, funded by EXC number 2064/1 – Project number 390727645. The authors would like to thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Thomas Klein.

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

imprs-is

**In a hurry?
Find this poster online!**