Don't trust your eyes: On the (un)reliability of feature visualizations

Robert **Geirhos***, Roland S. **Zimmermann***, Blair **Bilodeau***, Wieland **Brendel°**, Been **Kim°**

Google DeepMind

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

TL/DR How reliable are feature visualizations? We investigate this question through the lens of an adversary, empirically, and theoretically. All three perspectives cast doubt on the reliability of feature visualizations: They can be manipulated, don't reflect how natural input is processed & are provably unable to reliably predict even simple function behavior.

Motivation

Feature visualization is a **foundational interpretability** tool.

But are feature visualizations **reliable**, i.e. can we trust & rely on them?

We study this question from **three perspectives**:

Adversarial, Empirical, and heoretical.

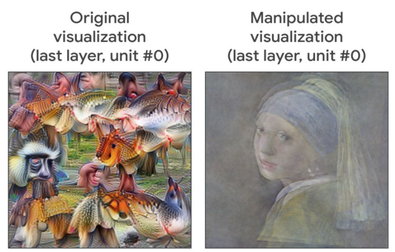
Summary

1. **Adversarial Perspective:** Feature visualizations can be fooled by manipulating a model.
2. **Empirical Perspective:** Even if the model is not manipulated, feature visualizations are processed largely along different paths compared to natural images, which means that they do not explain how neural networks process natural images.
3. **Theoretical Perspective:** Feature visualizations through activation maximization cannot be used to understand (i.e., predict the behavior of) black-box systems – instead, strong assumptions about the system are necessary.

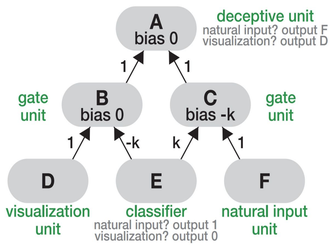
Conclusion: Feature visualization is best used for exploration / hypothesis generation, not for reliability or confirmation.
Potential way forward: Incorporate more structure & assumptions into networks (instead of seeking black-box explanations). More work needed!

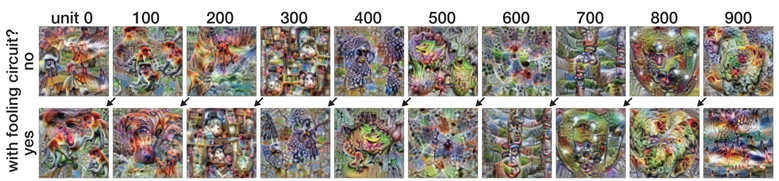
Adversarial Perspective

Through modifications through the network architecture, we can arbitrarily change feature visualizations while maintaining identical behavior on natural input.

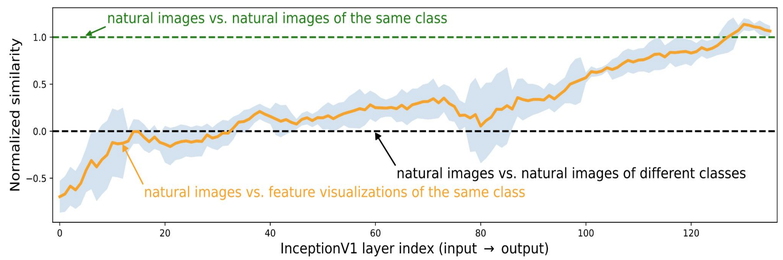


A "fooling circuit" manipulates visualizations:





Empirical Perspective



Intuition: Feature visualization aims to explain processing of natural input.

This is **only possible if visualizations are processed along similar paths** as natural input.

Empirically, that's not the case: Inception-V1 last-layer visualizations are processed along very different paths compared to natural input throughout the first 2/3rds of network layers.

Theoretical Perspective

Stronger assumptions about f ↓		Given feature visualization for f, can we reliably predict $f(x)$ ... exactly?	ε-approxim.?	closer to min or max?
black-box	$\mathcal{F}$	No	No	No
neural network (NN)	$\mathcal{F}_{\text{NN}}$	No	No	No
NN trained with ERM	$\mathcal{F}_{\text{ERM}}$	No	No	No
L–Lipschitz (known L)	$\mathcal{F}_{\text{Lip}}^{L}$	No	No	Only for small L
piecewise affine	$\mathcal{F}_{\text{PAff}}$	No	No	No
monotonic	$\mathcal{F}_{\text{Mono}}$	No	No	No
convex	$\mathcal{F}_{\text{Convx}}$	No	No	No
affine (input dim. > 1)	$\mathcal{F}_{\text{Aff}}^{d>1}$	No	No	No
affine (input dim. = 1)	$\mathcal{F}_{\text{Aff}}^{d=1}$	Yes	Yes	Yes
constant	$\mathcal{F}_{\text{Const}}$	Yes	Yes	Yes

Intuition: How much can you predict about a complex function from knowing its arg max? **Not much …**

We prove that **feature visualization** based on activation maximization **cannot be used to understand** (i.e., predict meaningful properties of) **a function** unless very strong additional knowledge about the function is available.

*RG, RSZ, BB: joint first authors (order between RSZ and BB determined by coin flip); WB & BK: joint senior authors (°). Affiliations: RG & BK: Google DeepMind; RSZ & WB: Max Planck Institute for Intelligent Systems & Tübingen AI Center; BB: University of Toronto, Department of Statistics