Scale Alone Does not Improve Mechanistic Interpretability in Vision Models

Roland Simon Zimmermann

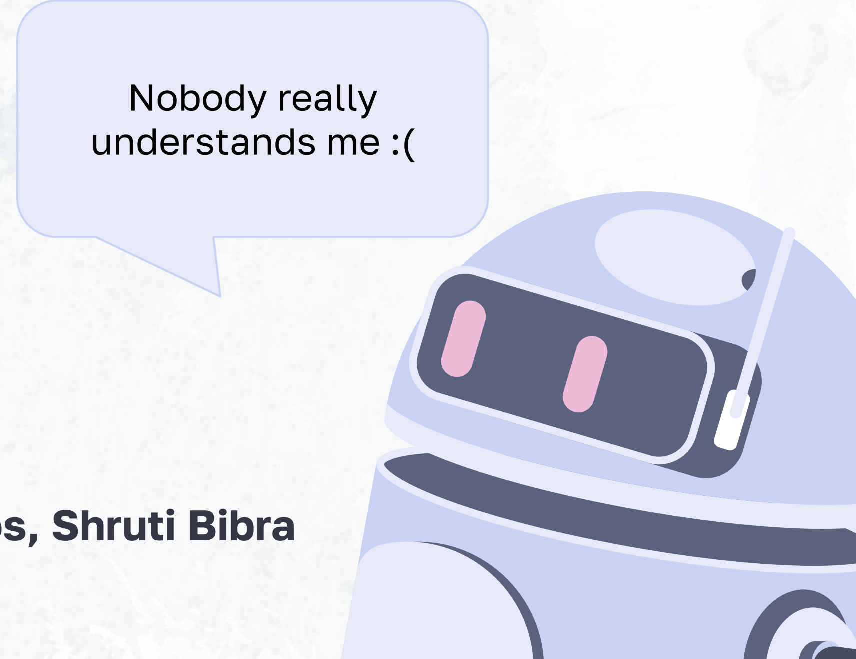
Google DeepMind
Verified email at google.com - Homepage
machine learning computer vision representation learning interpretability

Thomas Klein

PhD Student, University of Tübingen
Verified email at uni-tuebingen.de
Interpretability

Wieland Brendel

Fellow at ELLIS Institut Tübingen, Group Leader, Max Planck Institute for Intelligent Systems
Verified email at tuebingen.mpg.de - Homepage
machine learning computer vision

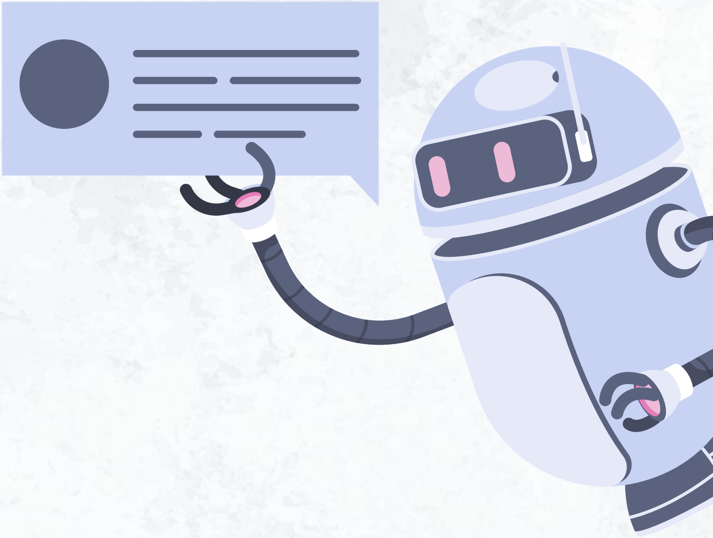


IFT6167

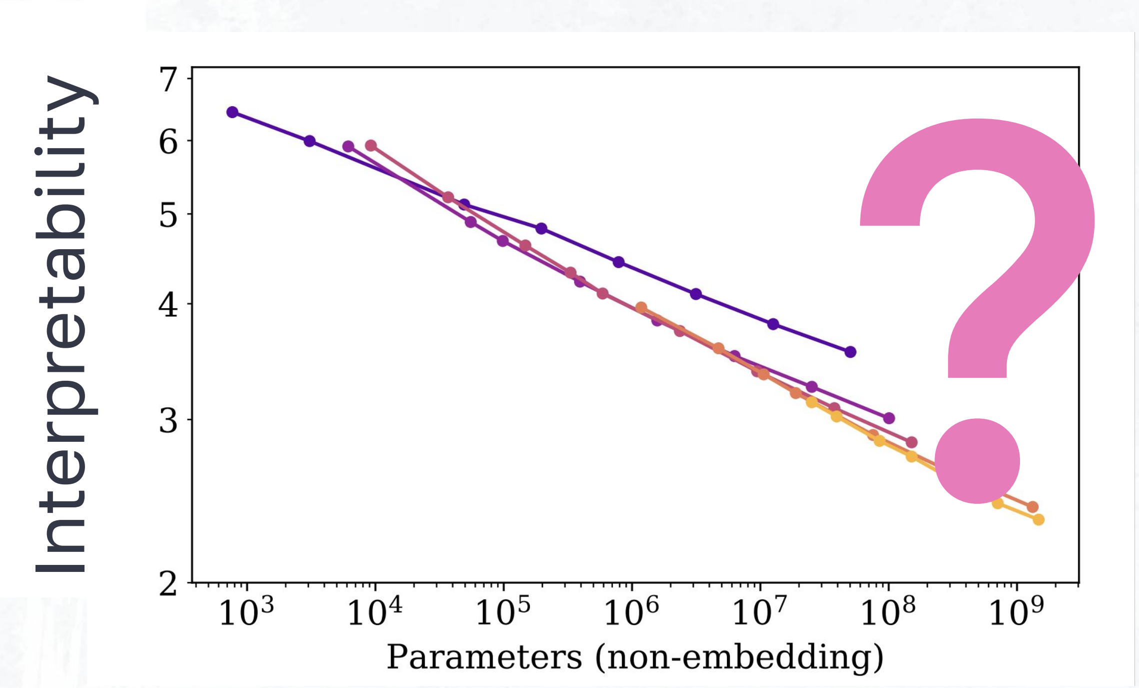
Presenters - Yorguin Jose Mantilla Ramos, Shruti Bibra

The question is...

What is the relation between the size of a model (e.g. scale) and its interpretability?

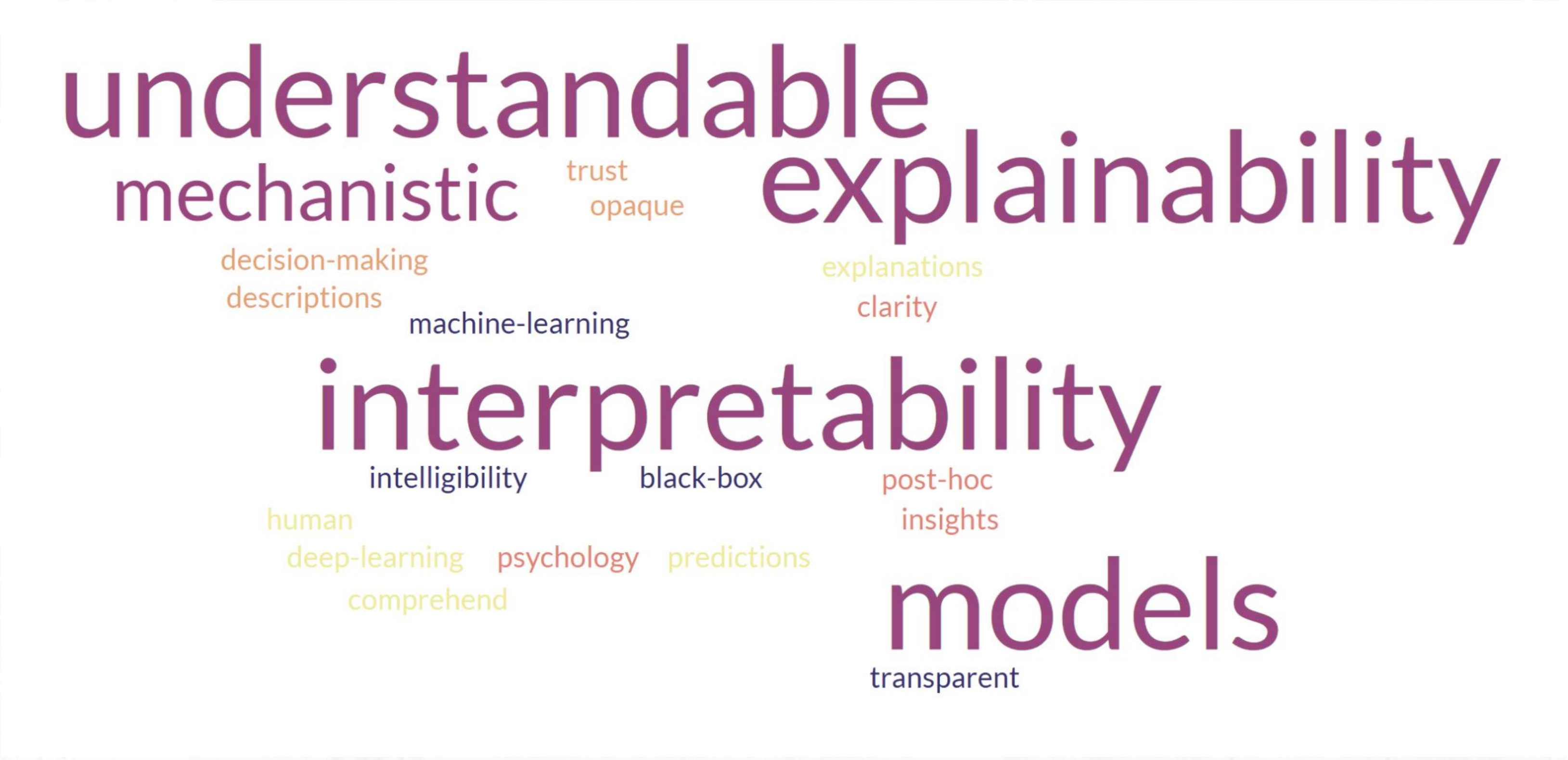


In other words...

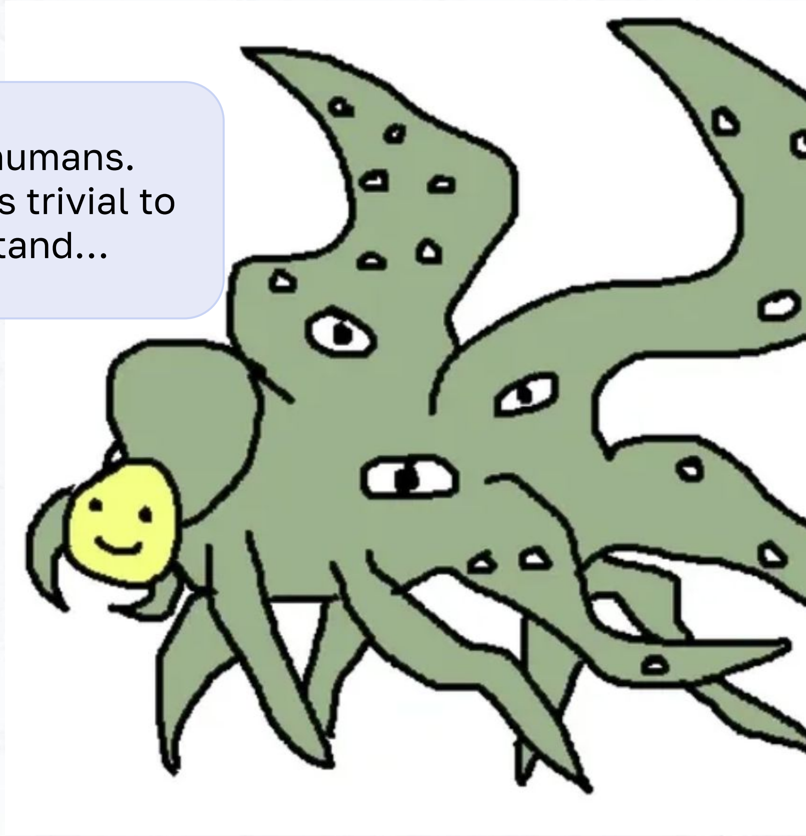


[1]

First, What is interpretability?



Ok… so how do you even measure it?
Isn't it a subjective property?



Psychophysics

E.g. : Magnitude estimation

Have the subject rate

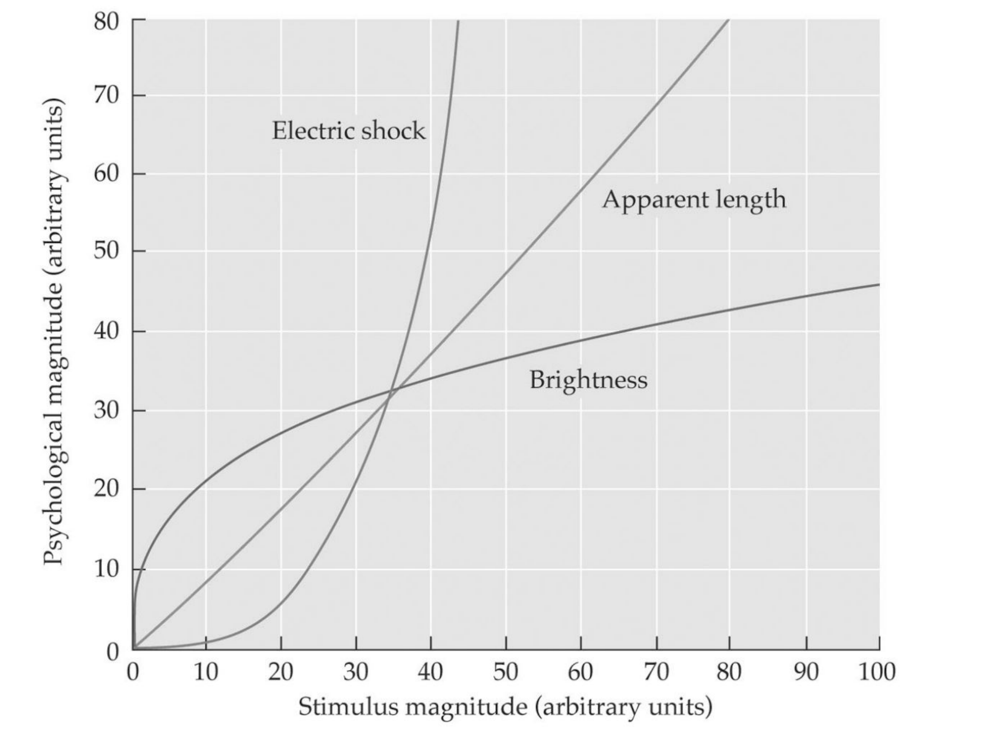
(e.g., 1-10)

some aspect of a stimulus

(e.g., how bright it appears or

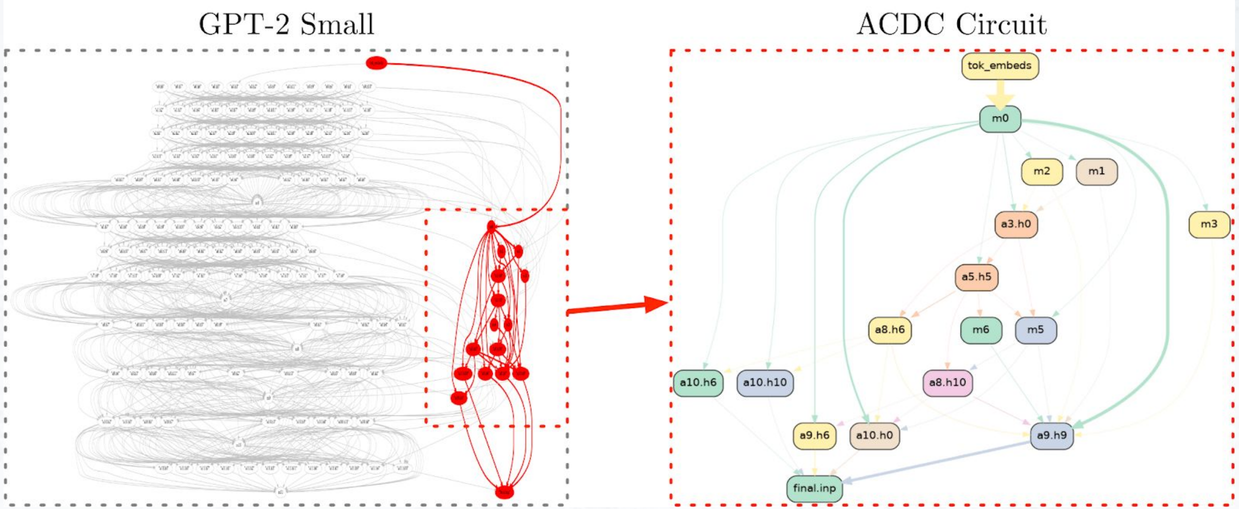
how loud it sounds) ..

[2]



SENSATION AND PERCEPTION, Figure 1.18 © 2006 Sinauer Associates, Inc.

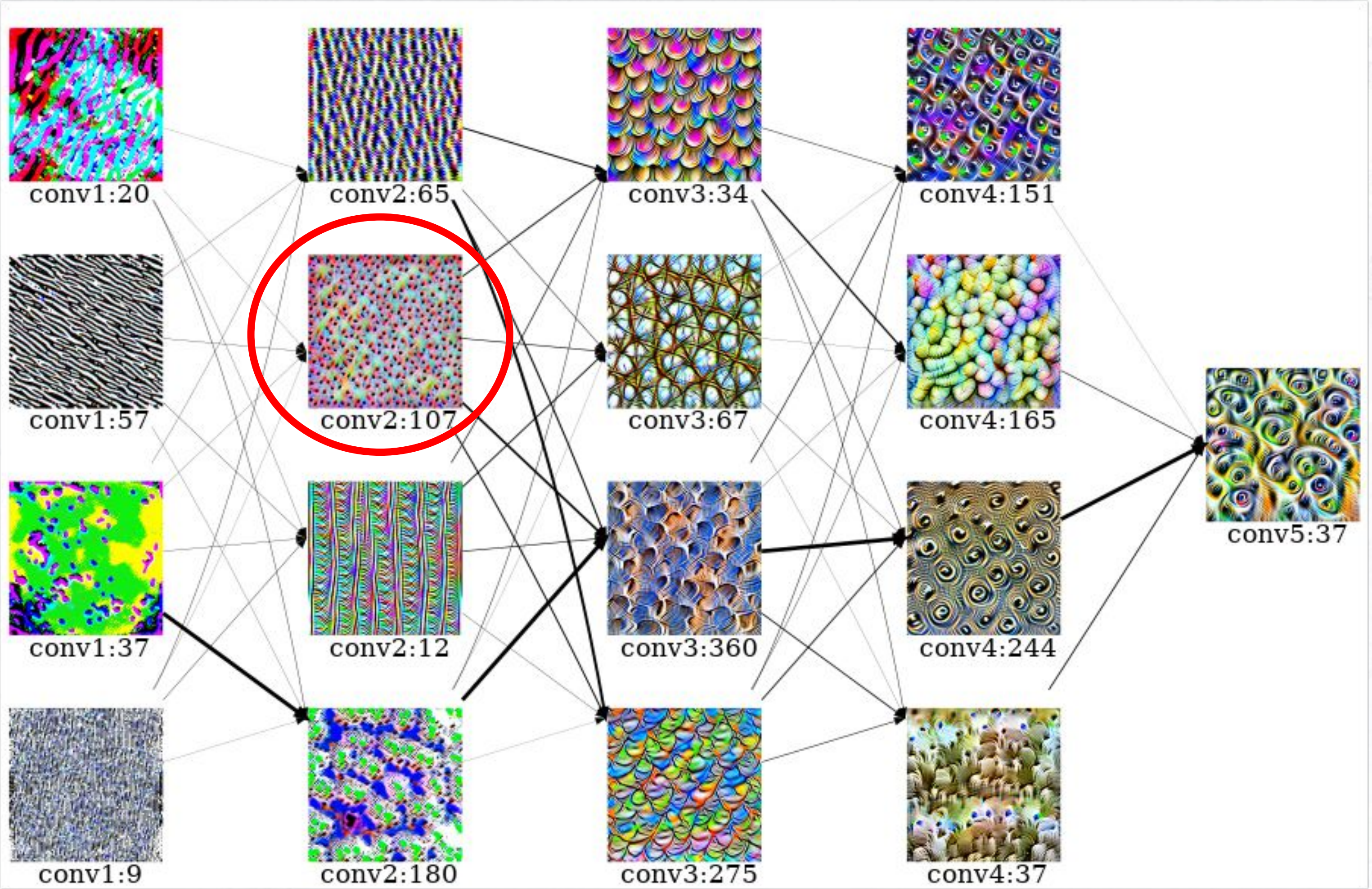
Still, interpretability seems a hard concept to measure in this way...



[3]

So, how can we framework interpretability as a question of perception?

Y


conv1:20
conv1:57
conv1:37
conv1:9
conv2:65
conv2:107
conv2:12
conv2:180
conv3:34
conv3:67
conv3:360
conv3:275
conv4:151
conv4:165
conv4:244
conv4:37
conv5:37

[4]

Natural and Synthetic Exemplars

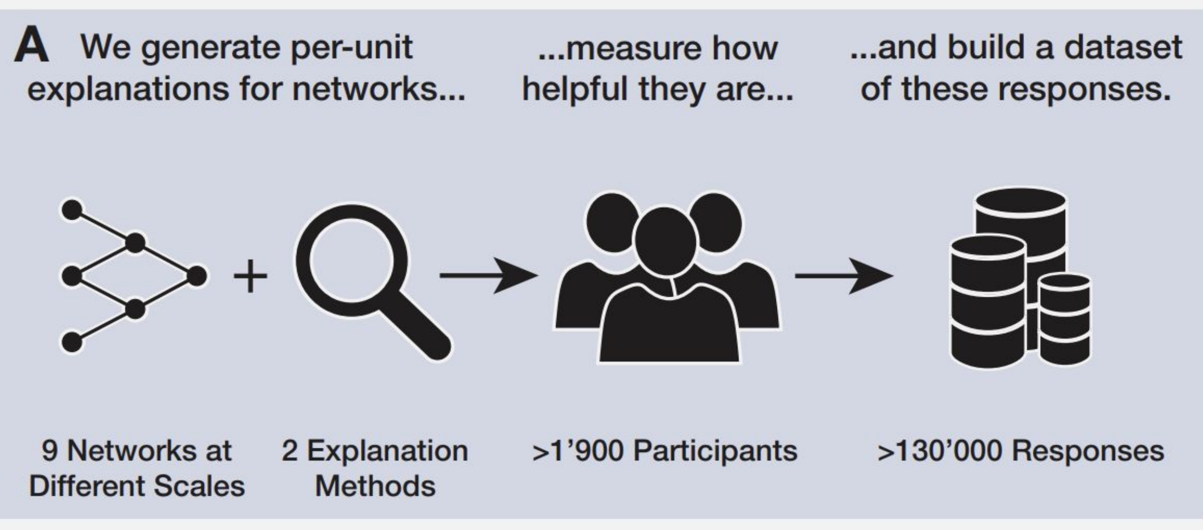
What are the advantages and disadvantages of each?

What was their hypothesis regarding the results?

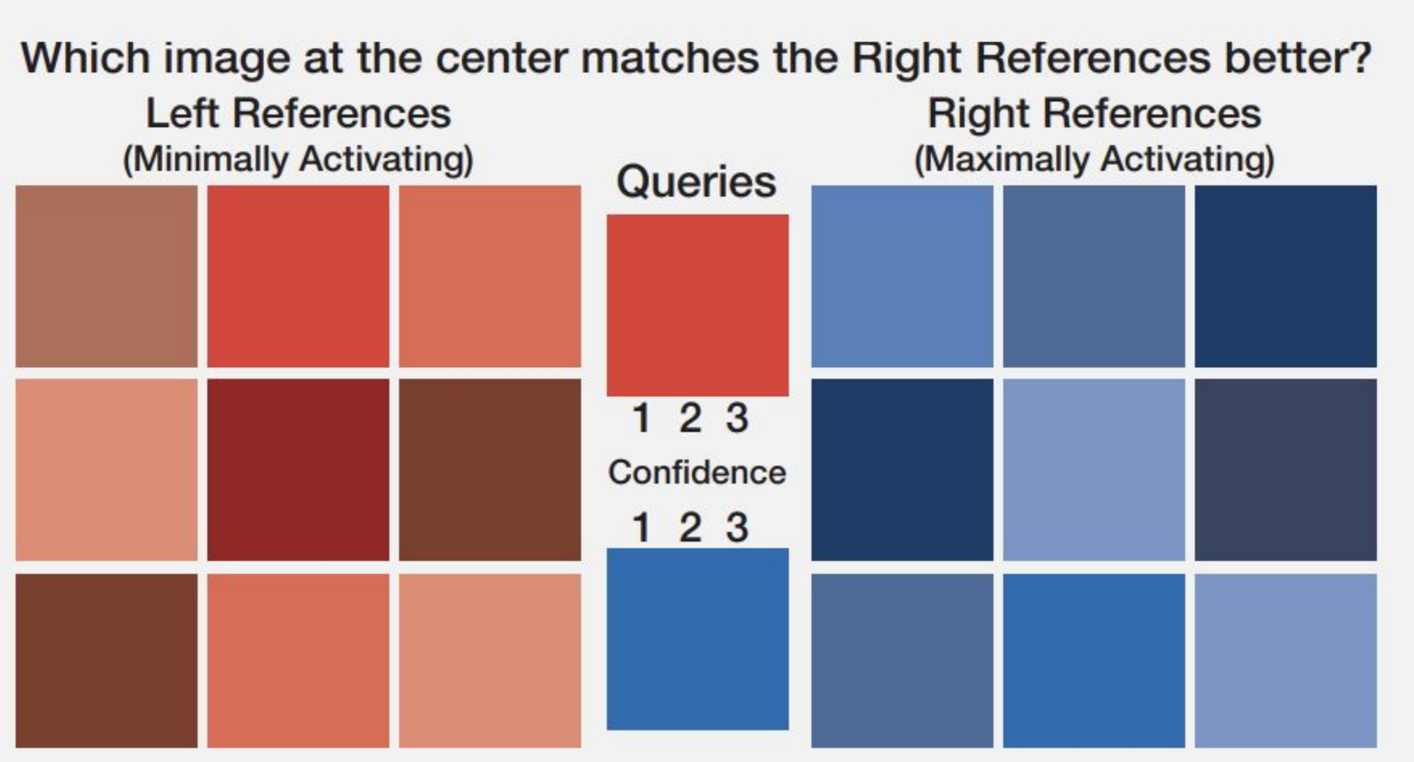
Scale Might Improve Per-Unit Interpretability

- Models trained on larger datasets align better with **human decision-making** (measured by error consistency).
  - **Possible reason**: Larger models may rely on **human-aligned, non-spurious features**, making their decisions more interpretable.
- Bigger models can **dedicate more units** to specific features.
  - This reduces **feature superposition**, making unit activations **less ambiguous** and easier to interpret.

Enough Context... What exactly was done?

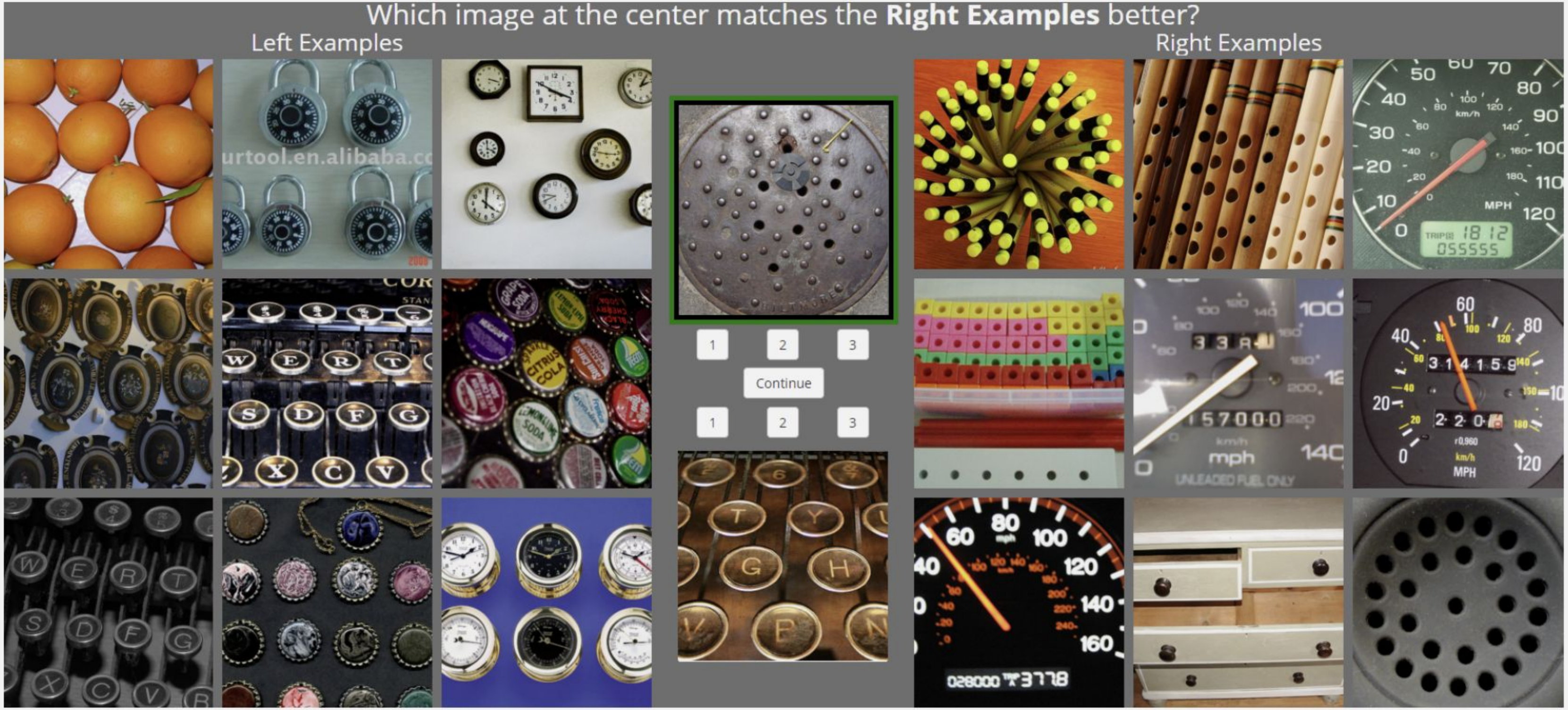


Task

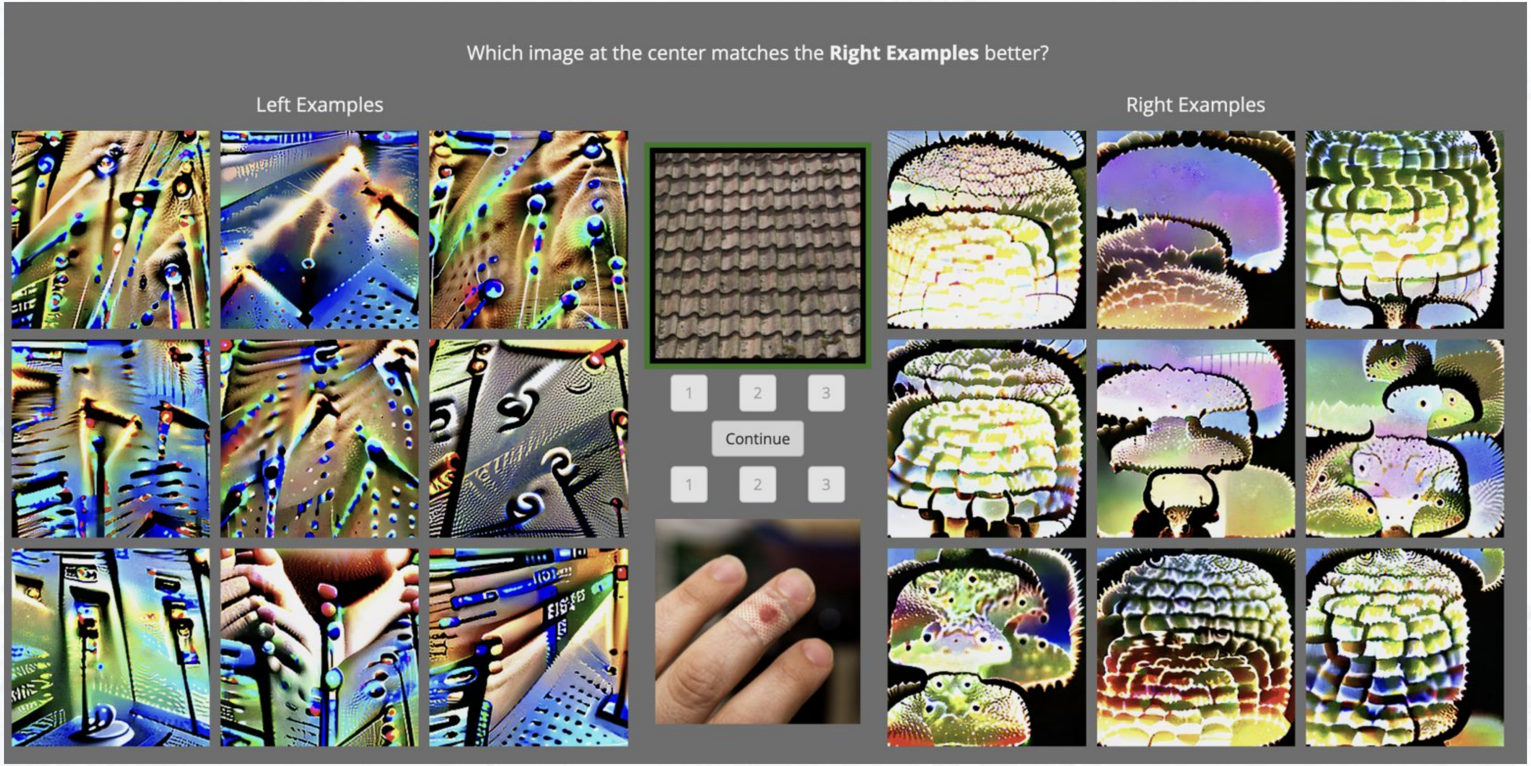

Which image at the center matches the Right References better?
Left References
(Minimally Activating)
Queries
Right References
(Maximally Activating)
1 2 3
Confidence
1 2 3

Y

Task (natural)

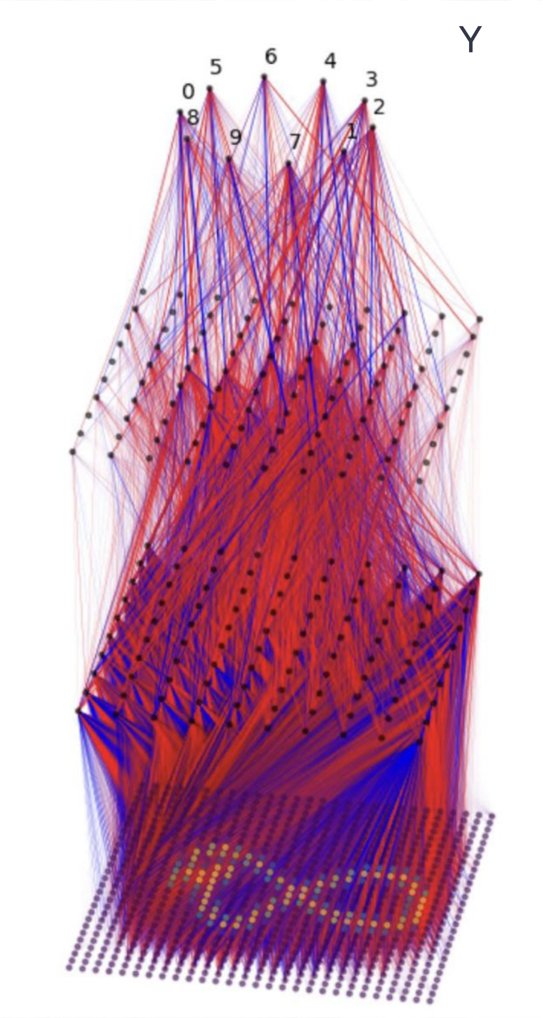


Task (synthetic)



for all the neurons???

- **Random Selection Process**
  - **84 units** selected per model.
  - First, **a network layer** is chosen from a **uniform distribution** over layers of interest.
  - Then, a **unit is randomly selected** from that layer.
- **Why Not Uniform Across All Units?**
  - CNNs **have more units in later layers**, so a simple uniform selection would bias toward them.
  - Instead, layers are **sampled first**, ensuring better representation across the model.
- **Layers of Interest**
  - **Convolution & normalization layers**.
  - **Outputs of skip connection blocks**.
  - **Exclusions**:
    - **First convolution layers** (can be analyzed directly via filters).
    - For **GoogLeNet**, only **last layers of inception blocks** are selected.
    - For **ViT models**, only **position-wise feedforward layers** are considered.



[5]

Y

Amazon Mechanical Turk

- **Geographic Restrictions**
  - Participants must be from **USA, Canada, UK, Australia, New Zealand, or Ireland**.
  - Ensures **English proficiency** and **ethical compensation**.
- **Experience & Reliability**
  - Must have completed **≥ 2,000 approved HITs**.
  - **Approval rate ≥ 99%** to ensure quality.
  - **No repeat participation** to prevent learning effects.
- **Attention & Engagement Filters**
  - **Demo trials**: Max **3 attempts** allowed.
  - **Reading time**: Must spend **≥ 15 seconds** on instructions.
  - **Catch trials**: Must answer **≥ 4 out of 5** correctly.
  - **Completion time**:
    - **Too fast**: < 135 seconds → **excluded**.
    - **Too slow**: > 2,500 seconds → **excluded**.
- **Behavioral Consistency**
  - Participants who **select the same query image > 90%** of the time are excluded.
- **Final Participant Selection**
  - **63 unique participants** per model pass quality checks.
  - Each participant completes **5 practice trials, 40 real trials, and 5 catch trials**.
  - **Total dataset**: **133,310 trials collected**, **76,000 valid trials retained**.
  - **Compensation**: $15/hour (~$2.79 per task).

Who participated?

With this setup you can dig into:

- **Interpretability vs scale / dataset size.**
- **Interpretability vs accuracy**
- **Interpretability vs human-likeness**
- **Interpretability vs Interpretability Methods (E.g. Natural vs Synthetic)**
  - **Interpretability vs Task Difficulty (e.g. selecting not the most/least activating examples).**
- **Interpretability vs Neuron-Unit Location**

Results

S

The 9 models tested and the varied dimensions

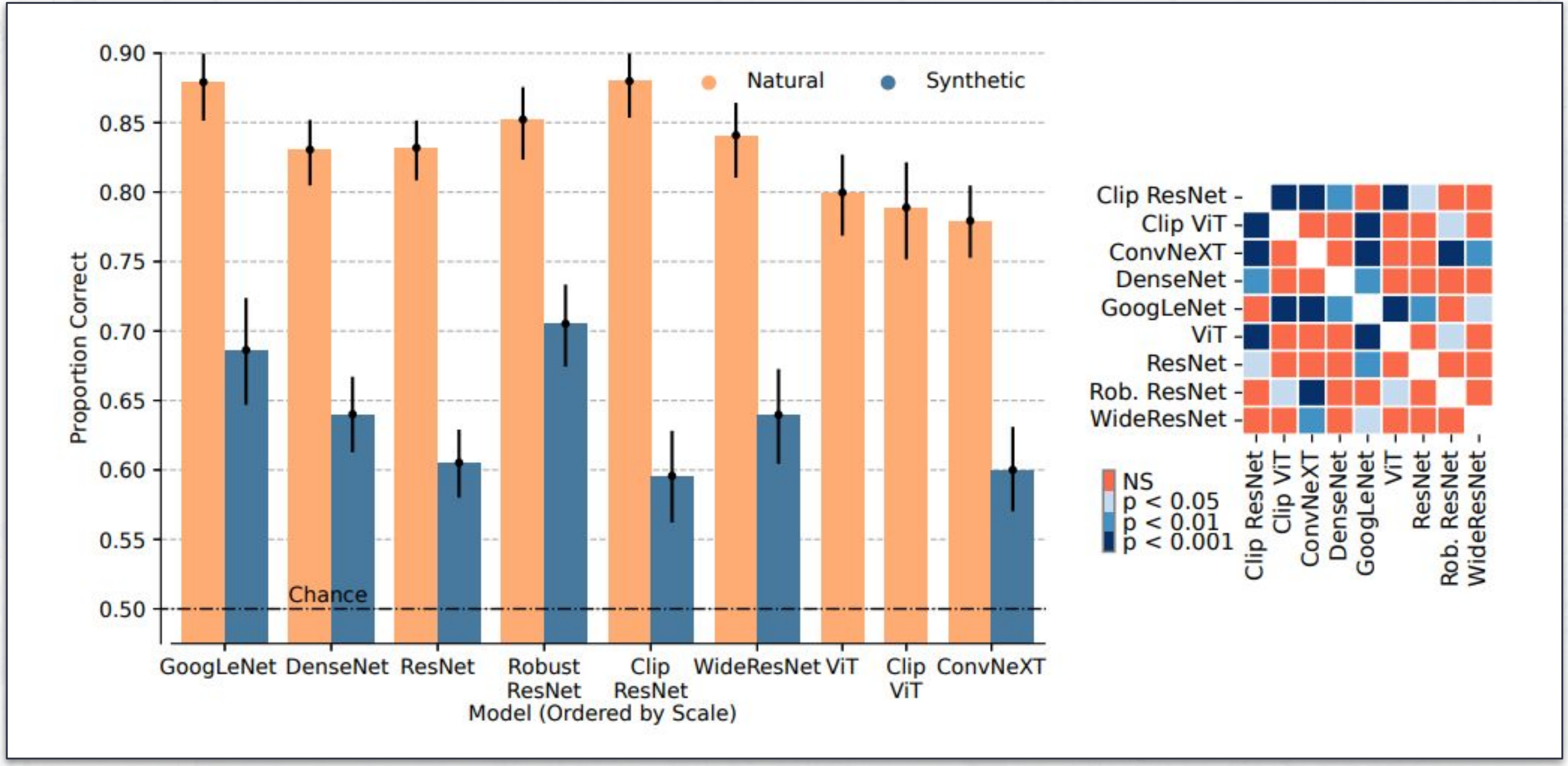
Model	Parameter Count	Design Aspect	Comparison Axis
GoogLeNet	6.8M	CNN, Inception modules	Baseline, Smallest Model
ResNet-50	25.6M	CNN, Residual connections	Deeper network
WideResNet-50	68.9M	CNN, Wider architecture	Increased model width
DenseNet-201	20.0M	CNN, Dense connectivity	Increased depth & connections
ViT-B	86M	Vision Transformer (ViT)	Transformer-based architecture
ConvNeXt-B	89M	CNN with modern improvements	Largest model
Clip ResNet-50	25.6M	CNN, Pretrained on LAION-400M	Large-scale dataset training
Clip ViT-B	86M	Vision Transformer, Pretrained	Large-scale dataset training
Robust ResNet-50	25.6M	CNN, Adversarial robustness	Tested for robustness & interpretability

S

Q1. Does scaling models improve interpretability ?

Q1. Does scaling models improve interpretability ?

NO!



Natural
Synthetic
Proportion Correct
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Chance
GoogLeNet
DenseNet
ResNet
Robust ResNet
Clip ResNet
WideResNet
ViT
Clip ViT
ConvNeXT
Model (Ordered by Scale)
Clip ResNet
Clip ViT
ConvNeXT
DenseNet
GoogLeNet
ViT
ResNet
Rob. ResNet
WideResNet
NS
p < 0.05
p < 0.01
p < 0.001

Q1. Does scaling models improve interpretability ?

NO!

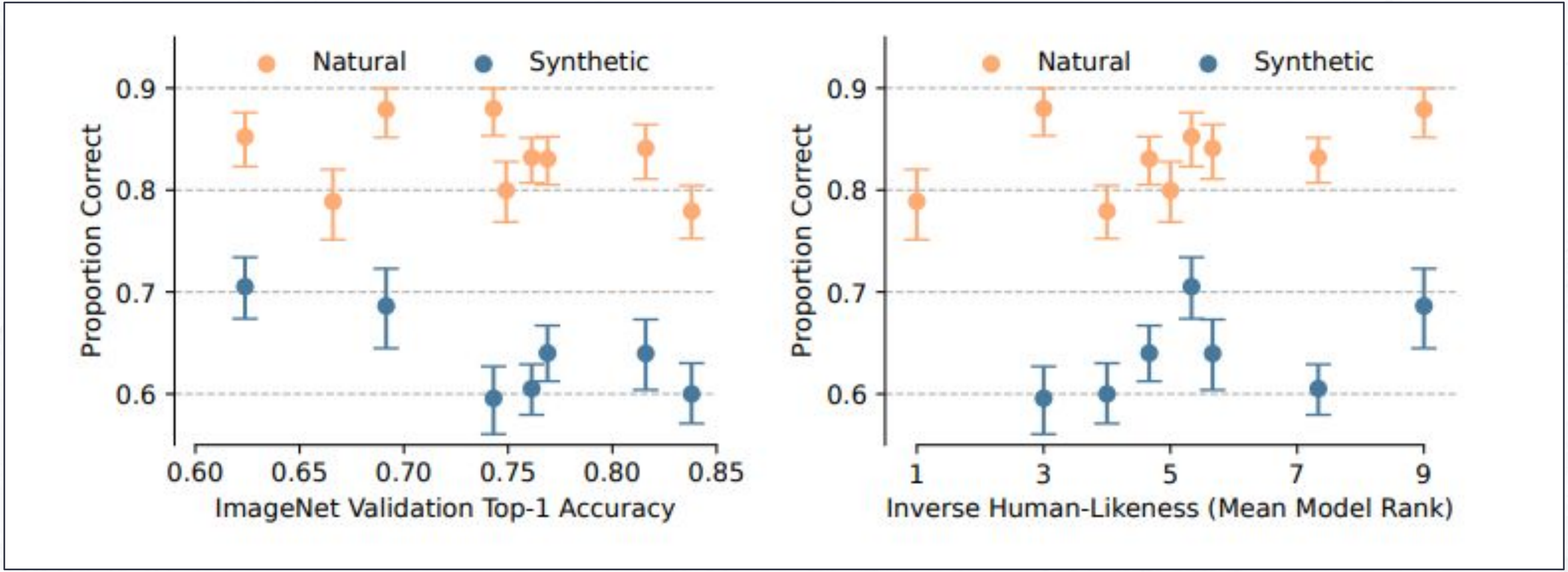
1. Clearly! No improvement
2. GoogleNet model performs way better than Vit

Q2. Does higher classification performance or human-like decisions translate to high mechanistic interpretability?

Q2. Does higher classification performance or human-like decisions translate to high mechanistic interpretability?

NO!

S



Natural
Synthetic
0.9
0.8
0.7
0.6
Proportion Correct
0.60
0.65
0.70
0.75
0.80
0.85
ImageNet Validation Top-1 Accuracy
Natural
Synthetic
0.9
0.8
0.7
0.6
Proportion Correct
1
3
5
7
9
Inverse Human-Likeness (Mean Model Rank)

Q2. Does higher classification performance or human-like decisions translate to high mechanistic interpretability?

NO!

1. As human likeness increase, interpretability decreases(for synthetic)
2. No positive relationship

Q3. Is synthetic feature visualization technique helpful ?

Q3. Is synthetic feature visualization technique helpful ?

NO!

Q3. Is synthetic feature visualization technique helpful ?

NO!

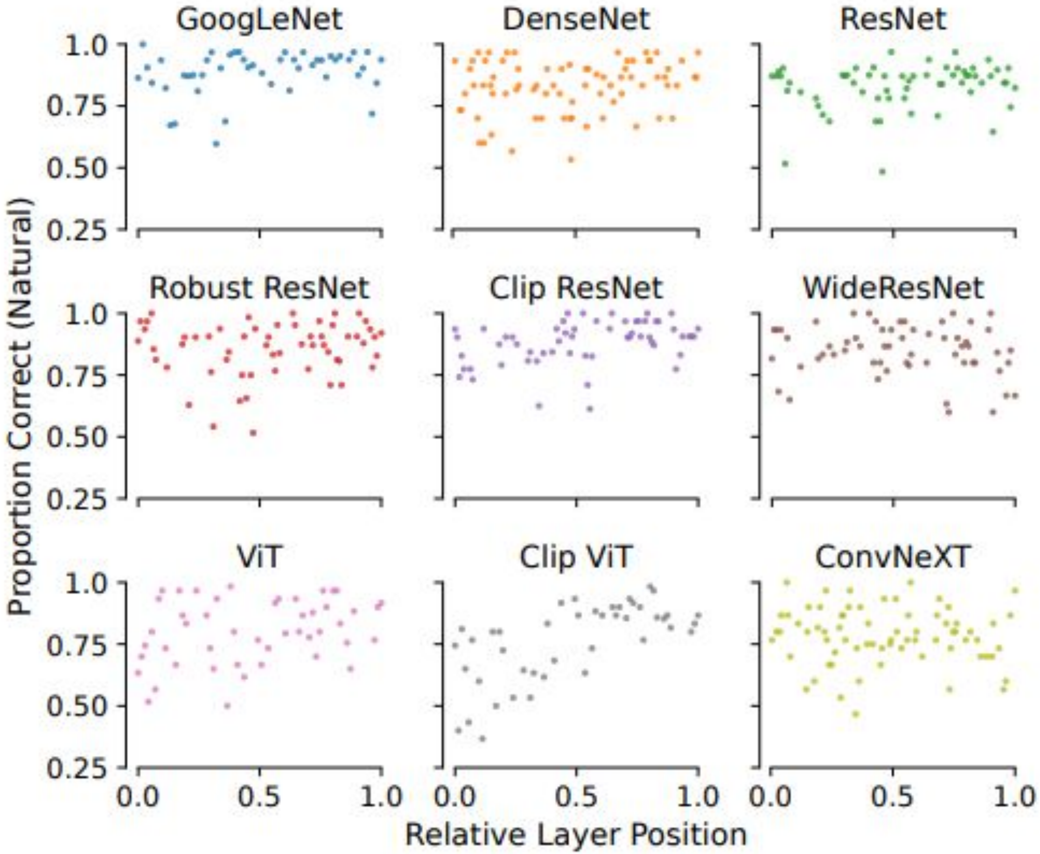
1. Evidence from the previous graphs!
2. Natural exemplars are better for understanding neuron behaviour

Q4. Is any specific layer a stronger predictor of interpretability?

S

Q4. Is any specific layer a stronger predictor of interpretability?

Somewhat!



Model	r
GoogleNet	.28
DenseNet	.17
ResNet	.16
Rob. ResNet	.09
Clip ResNet	.39**
WideResNet	−.16
ConvNeXT	−.09
ViT	.26
Clip ViT	.64***

S

Q4. Is any specific layer a stronger predictor of interpretability?

Somewhat!

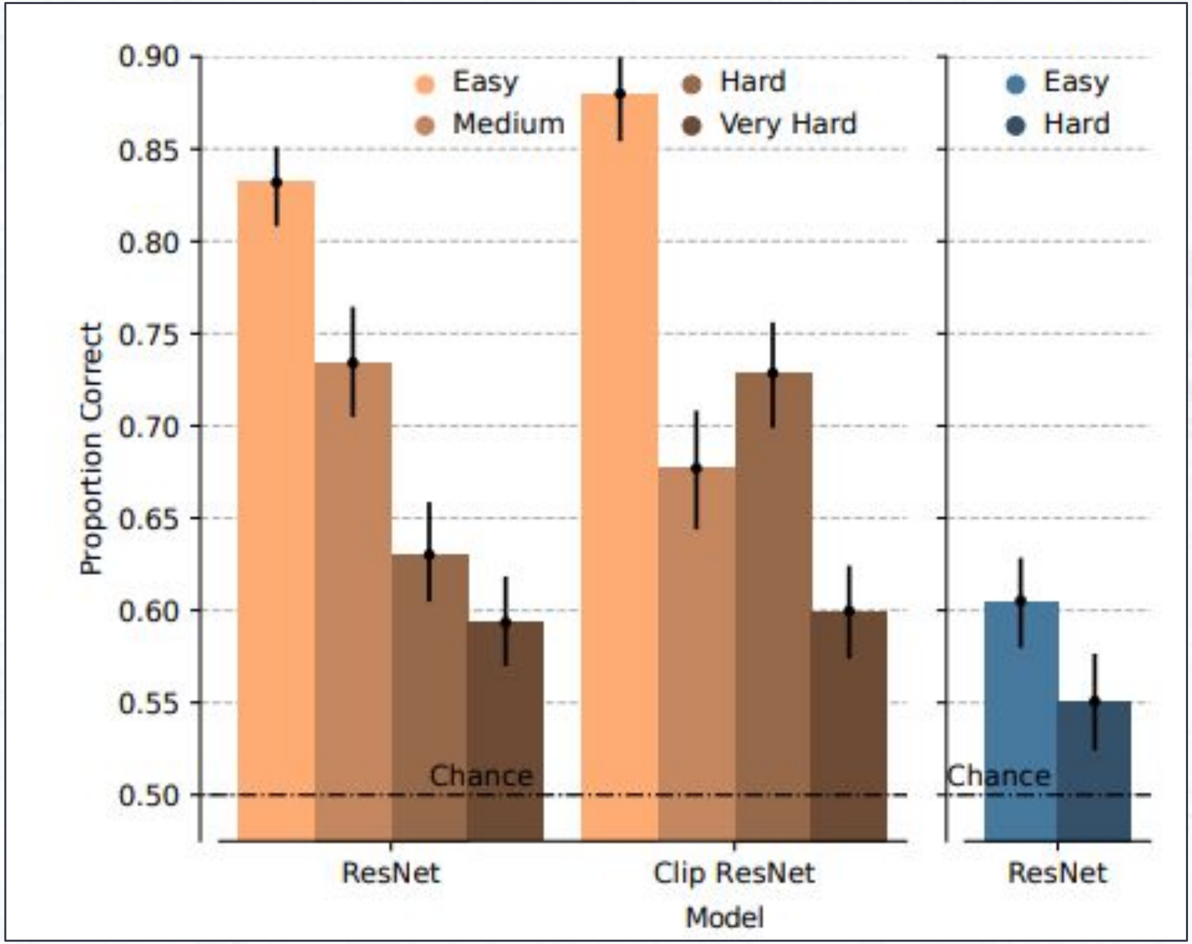
1. Consistent correlation
2. ClipResnet and ClipVit - more interpretability in later layers

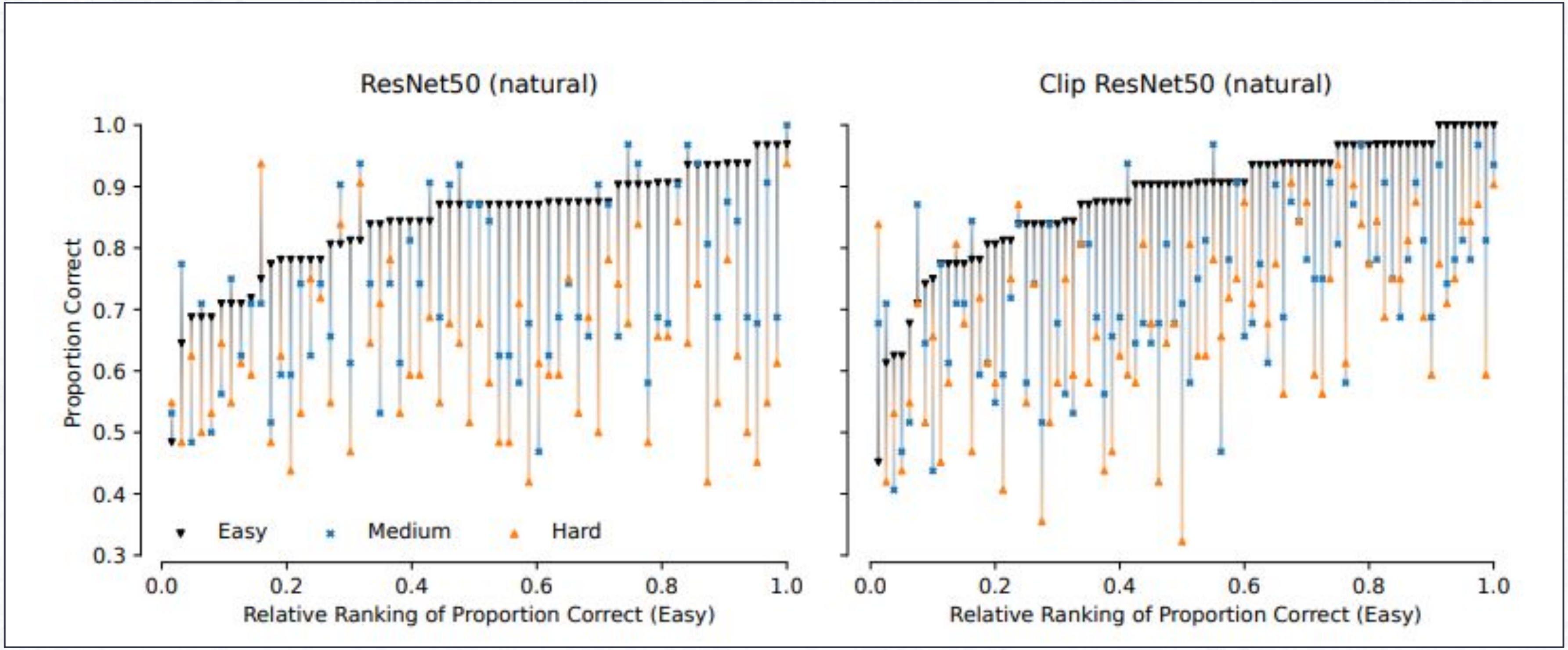
Q5. Does task difficulty affects interpretability?

Q5. Does task difficulty affects interpretability?

Yes!

S


Easy
Medium
Hard
Very Hard
Easy
Hard
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
Proportion Correct
Chance
Chance
ResNet
Clip ResNet
ResNet
Model


ResNet50 (natural)
Clip ResNet50 (natural)
Proportion Correct
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
Easy
Medium
Hard
0.0
0.2
0.4
0.6
0.8
1.0
Relative Ranking of Proportion Correct (Easy)
Relative Ranking of Proportion Correct (Easy)

Q5. Does task difficulty affects interpretability?

Yes!

Interpretability decreases with task difficulty

S

Disappointing results!!!

But, what did we gain from this study ?

IMI - A Dataset to Learn Automated Interpretability Measures

Name	Size
human_responses.zip md5:f886fc48a87baf51f2beb834924c8b62	61.9 MB
image_data.zip md5:47c364fd92752d3412f1c08f8cd6d793	1.6 GB

https://zenodo.org/records/8131197

1. Models need to be explicitly optimized for interpretability
2. Enable research on automated interpretability measures
3. 130,000 anonymized human responses, each containing a final choice, confidence score, and reaction time
4. 76,000 of the responses passed quality checks The dataset also includes the query images and the generated explanations for 767 units across nine different models.

Conclusion

No practically relevant differences...

As our study shows, new model design choices or training objectives are needed to *explici*
the mechanistic interpretability of vision models. We expect the data collected in our study

- Results are not surprising?
  - Accuracy/Interpretability trade-off (∃ debates about it).
  - But the paper claim is that there is no relation.
- Is this a true estimation of “interpretability”?
  - Visualizations (Natural and Synthetic) has been criticized (unreliable, misleading) in the literature, including the authors .
  - Feature Visualizations do not sufficiently explain hidden units of Artificial Neural Networks
- Results could benefit from more controlled experiments.
  - E.g. manipulating model and dataset size in a single model and training it. (↑ comp.)

References

[0] R. S. Zimmermann, T. Klein, and W. Brendel, “Scale Alone Does not Improve Mechanistic Interpretability in Vision Models,” 2023, arXiv. doi: 10.48550/ARXIV.2307.05471.
[1] J. Kaplan et al., “Scaling Laws for Neural Language Models,” 2020, arXiv. doi: 10.48550/ARXIV.2001.08361.
[2] Wolfe, J. M., Kluender, K. R., Levi, D. M., Bartoshuk, L. M., Herz, R. S., Klatzky, R. L., & Lederman, S. J. (2006). Sensation and perception. Sinauer Associates.
[3] A. Conmy, A. N. Mavor-Parker, A. Lynch, S. Heimersheim, and A. Garriga-Alonso, “Towards Automated Circuit Discovery for Mechanistic Interpretability,” 2023, arXiv. doi: 10.48550/ARXIV.2304.14997.
[4] G. Nanfack, M. Eickenberg, and E. Belilovsky, “From Feature Visualization to Visual Circuits: Effect of Adversarial Model Manipulation,” 2024, arXiv. doi: 10.48550/ARXIV.2406.01365.
[5] Z. Liu, E. Gan, and M. Tegmark, “Seeing is Believing: Brain-Inspired Modular Training for Mechanistic Interpretability,” 2023, arXiv. doi: 10.48550/ARXIV.2305.08746.

Thank you!