Reverse Engineering the visual system

PI: Thomas Serre
Victor Boutin
16/11/2023
Artificial vision = Biological vision?

Claimed to be «human level»

SOURCE: PAPERSWITHCODE.COM
Artificial vision = Biological vision?

Prediction: Cat

SZEGEDY ET AL 2013
Artificial vision = Biological vision?

**Prediction**: Cat

+ 0.25 x
Artificial vision = Biological vision?

Prediction: Cat

Prediction: Ostrich

\[
\text{Prediction} : \text{Cat} + 0.25 \times \text{Ostrich} = \text{Cat}
\]

SZEGEDY ET AL 2013
Reverse Engineering the Visual System
Reverse Engineering the Visual System

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms
Reverse Engineering the Visual System

• Benchmark for compositional visual reasoning

Zerroug et al 2022

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms
Reverse Engineering the Visual System

• Benchmark for compositional visual reasoning
  Zerroug et al. 2022

• Neuro plausible attentional mechanisms
  Vaishnav et al. 2022
Reverse Engineering the Visual System

- Benchmark for compositional visual reasoning
  Zerroug et al. 2022

- Neuro plausible attentional mechanisms
  Vaishnav et al. 2022

- Leveraging binding by synchrony in complex networks (with VanRullen)
  Muzellec et al. 2023
Reverse Engineering the Visual System

We test « humanness » of AI using XAI and metrics from cognitive science

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms
Reverse Engineering the Visual System

• Improving XAI methods


We test « humanness » of AI using XAI and metrics from cognitive science

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms
We test « humanness » of AI using XAI and metrics from cognitive science

• Improving XAI methods

(Fei et al. 2021, Fei et al. 2022A, Fei et al. 2022B, Fei et al. 2023A, Fei et al. 2023B)

• https://serre-lab.github.io/Lens/

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms

Reverse Engineering the Visual System
We test «humanness» of AI using XAI and metrics from cognitive science.

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms.

- Improving XAI methods (FELMETAL 2021, FELMETAL 2022A, FELMETAL 2022B, FELMETAL 2023A, FELMETAL 2023B)
- https://serre-lab.github.io/Lens/

Reverse Engineering the Visual System

Concept 1 (35%)
Concept 2 (23%)
Concept 3 (20%)
We test « humanness » of AI using XAI and metrics from cognitive science.

- Improving XAI methods
  (FEL et AL 2021, FEL et AL 2022A, FEL et AL 2022B, FEL et AL 2023A, FEL et AL 2023B)

- Harmonizing machines and humans with XAI
  (FEL et AL 2022C)

Train AI on tasks inspired by cognitive science to highlight key computational mechanisms.

https://serre-lab.github.io/Lens/
Reverse Engineering the Visual System

We test « humanness » of AI using XAI and metrics from cognitive science.

The chair exemplified with the one-shot drawing project …

Train AI on tasks inspired by neuroscience to highlight key computational mechanisms.
One-Shot Drawing Task (Lake et al. 2015)
One-Shot Drawing Task (Lake et al. 2015)
One-Shot Drawing Task (Lake et al. 2015)

Training

Exemplars

Variations
One-Shot Drawing Task (Lake et al. 2015)

Training

Exemplars → Variations

Testing

New exemplars → Variations
One-Shot Drawing Task (Lake et al. 2015)

Omniglot (Lake et al. 2015)

Quick, Draw! (Ha et al. 2017)
Task Evaluation: Originality vs Recognizability (BOUTIN ET AL 2022)
Task Evaluation: Originality vs Recognizability (BOUTIN ET AL 2022)

- Recognizability (classification accuracy)
- Originality (Mean distance to exemplar)

Evaluated using a one-shot classifier

\[ \| f(\text{Exemplar}) - f(\text{Variation}) \|_2 \]
Task Evaluation: Originality vs Recognizability (BOUTIN ET AL 2022)

Recognizability (classification accuracy)

Originality (Mean distance to exemplar)

- Recognizability
- Originality

- Good generalization

Evaluated using a one-shot classifier

Generated samples

Exemplar

Classifier decision boundary

\[ \| f(\text{Exemplar}) - f(\text{Variation}) \|_2 \]
Task Evaluation: Originality vs Recognizability (Boutin et al. 2022)

- Recognizability (classification accuracy)
- Originality (Mean distance to exemplar)

Good generalization

Bad generalization

Evaluated using a one-shot classifier

Classifier decision boundary

Generated samples

Exemplar

Exemplar

Variation

\( f(\text{\textbullet}) - f(\text{\textbullet}) \)
Models in the Originality vs. Recognizability Space (Boutin et al. 2023)

Omniglot

Quick, Draw!

Recallability (%) vs. Mean Originality (Normalized)
Models in the Originality vs. Recognizability Space (Boutin et al. 2023)

Omniglot

Quick, Draw!
Models in the Originality vs. Recognizability Space (Boutin et al. 2023)

Omniglot

Quick, Draw!

[Graphs showing the distribution of models in the originality vs. recognizability space for Omniglot and Quick, Draw!]
Models in the Originality vs. Recognizability Space (Boutin et al. 2023)

VAE over-generalizes

GAN drops modes

Data

(GAN) (Lucas et al. 2019)
Models in the Originality vs. Recognizability Space (Boutin et al, 2023)

Omniglot

Quick, Draw!

Recognizability (%) vs. Mean Originality (Normalized)

- DA-GAN-UN
- DA-GAN-RN
- VAE-STN
- VAE-NS
- CFGDM
- DDPM
- FSDM
- Human

Recognizability (%) vs. Mean Originality (Normalized)

- DA-GAN-UN
- DA-GAN-RN
- VAE-STN
- VAE-NS
- CFGDM
- DDPM
- FSDM
- Human
Models in the Originality vs. Recognizability Space (BOUTIN ET AL 2023)

Omniglot

Quick, Draw!

[Graphs showing data distribution in the Originality vs. Recognizability space for Omniglot and Quick, Draw!]
Can you tell apart human from machine-generated samples?
Can you tell apart human from machine-generated samples?
Generalization Curves: Recognizability = f(Originality)
Generalization Curves: Recognizability = f(Originality)
Important Features for Humans and Machines
Important Features for Humans and Machines

• Human importance maps collected using the ClickMe challenge (LINSLEY ET AL 2019)
Important Features for Humans and Machines

• Human importance maps collected using the ClickMe challenge (LINSLEY ET AL 2019)
Important Features for Humans and Machines

- Human importance maps collected using the ClickMe challenge *(LINSLEY ET AL 2019)*

- **CFGDM** importance maps using attribution methods
Important Features for Humans and Machines
Conclusion/Discussion

Originality vs Recognisability to evaluate generation performance of humans and machines
Conclusion/Discussion

Originality vs Recognisability to evaluate generation performance of humans and machines

Diffusion models fail at modelling original drawings
Conclusion/Discussion

Originality vs Recognisability to evaluate generation performance of humans and machines

Diffusion models fail at modelling original drawings

Different attentional strategies leveraged by humans and machines
Thank you for your attention …

2023
• A holistic approach to unifying automatic concept extraction and concept importance estimation (Neurips 2023), T. Fel, V. Boutin, M. Moayeri, R. Cadene, L. Bethune, L. Andeol, M. Chalvidal & T. Serre
• Learning functional transduction (Neurips 2023), M. Chalvidal, T. Serre & R. VanRullen
• GAMR: A Guided Attention Model for (visual) Reasoning (ICLR 2023), M. Vaishnav & T. Serre

2022
• Harmonizing the object recognition strategies of deep neural networks with humans (Neurips 2022), T. Fel*, I.F. Rodriguez*, D. Linsley* & T. Serre
• A benchmark for compositional visual reasoning (Neurips 2022), A. Zerroug, M. Vaishnav, J. Colin, S. Musslick & T. Serre
• Diversity vs. recognizability: Human-like generalization in one-shot generative models (Neurips 2022), V. Boutin, L. Singhal, X. Thomas & T. Serre
• Meta-reinforcement learning with self-modifying networks (Neurips 2022), M. Chalvidal, T. Serre, R. VanRullen
• Understanding the computational demands underlying visual reasoning, (Neural Computation), M. Vaishnav, R. Cadene, A. Alamia, D. Linsley, R. VanRullen & T. Serre

2021
• Go with the flow: Adaptive control for Neural ODEs (ICLR 2021), M. Chalvidal, M. Ricci, R. VanRullen, T. Serre
• Iterative VAE as a predictive brain model for out-of-distribution generalization (Neurips workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM)), V. Boutin, A. Zerroug, M. Jung, & Thomas Serre