Artificial Cognition

Shashank Shekhar
Masters in Applied Science candidate,
Machine Learning Research Group, University of Guelph
Vector Research Scholar, Vector Institute

VECTOR INSTITUTE | UNIVERSITY of GUELPH

Credits

Eric J Taylor
Postdoc, Vector Institute
(slides borrowed almost in their entirety from Eric)

Graham Taylor
University of Guelph, Vector Institute, Canada CIFAR AI Chair

Disclaimer

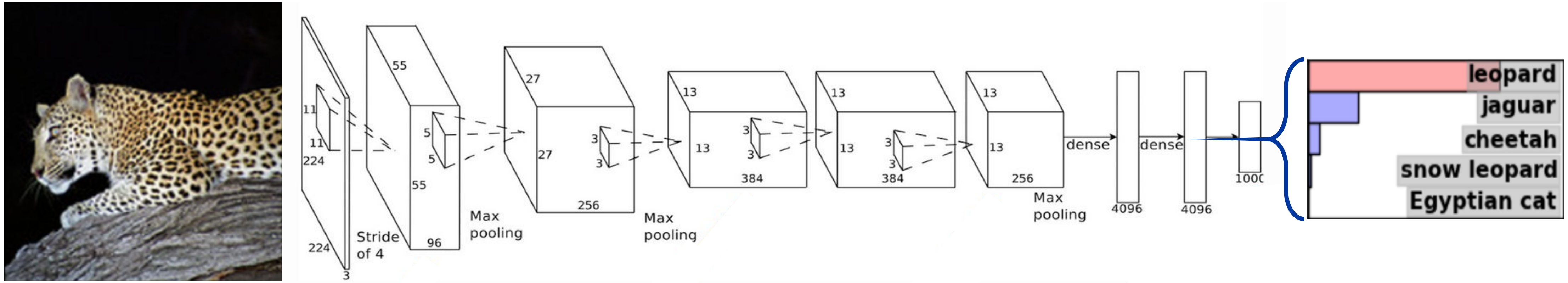
- Assumes that the audience has knowledge at an introductory machine learning course level
- Machine learning <-> Deep learning <-> Computer vision (in the context of this presentation)
- The presentation is more of an overview of the field than any sort of deep dive

Table of Contents

Table of Contents

Explainable AI

The Black Box Problem

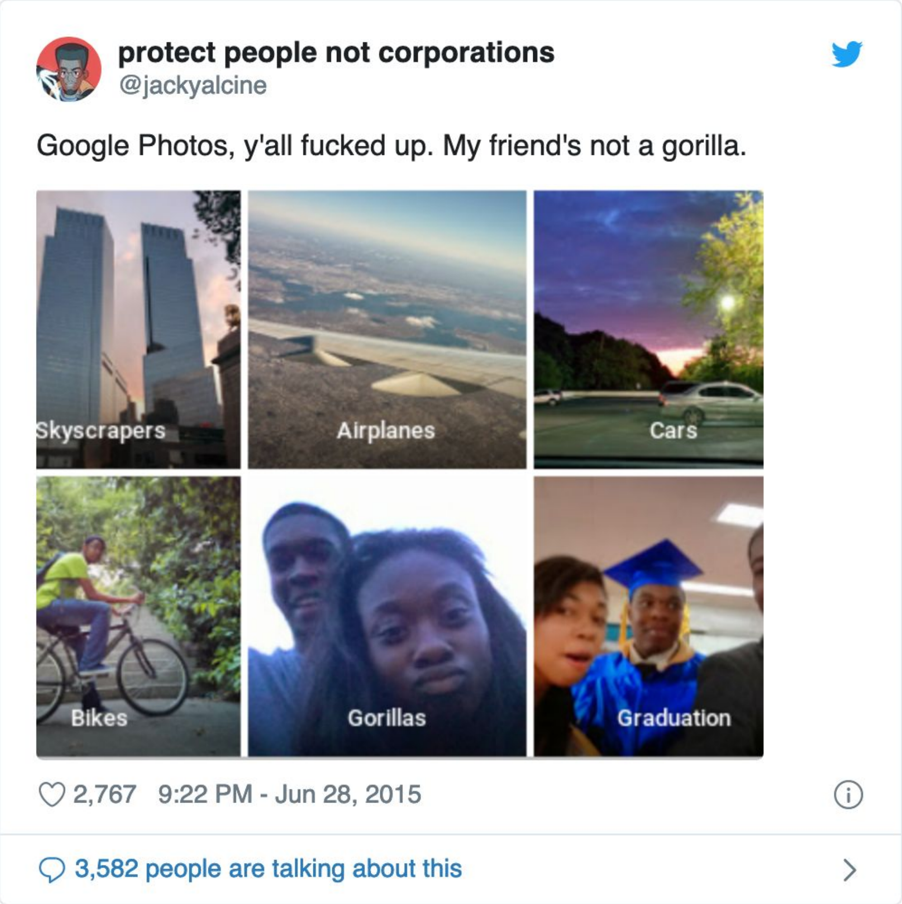


- Contains enough operations that you would not be able to compute a forward pass in your lifetime
- This (outdated) model contains 772 840 neurons; how can we make sense of it?
- Called a black-box problem because the decision appears inscrutable from the outside
- Even if we examine the code, trained parameters, or elementary operations, it is difficult or impossible to express how they combine to form a decision

(1) Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).
(2) Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network?. arXiv preprint arXiv:1907.06374.

Explainable AI

Ethical & Political Impetus



- Challenges in XAI highlight the need for good explanations for why and how deep learning algorithms make decisions
- Why do image classification algorithms make racist choices? Under what conditions might an autonomous vehicle hit a pedestrian? When are we confident enough in an assisted medical diagnosis to use it in field?
- Data protection laws in Europe give citizens a “right to an explanation” when an algorithm makes a decision that affects them
- France may require the communication of model parameters
- Google’s solution to the racist classifier was not to explain the decision but to remove gorillas from the list of possible classes

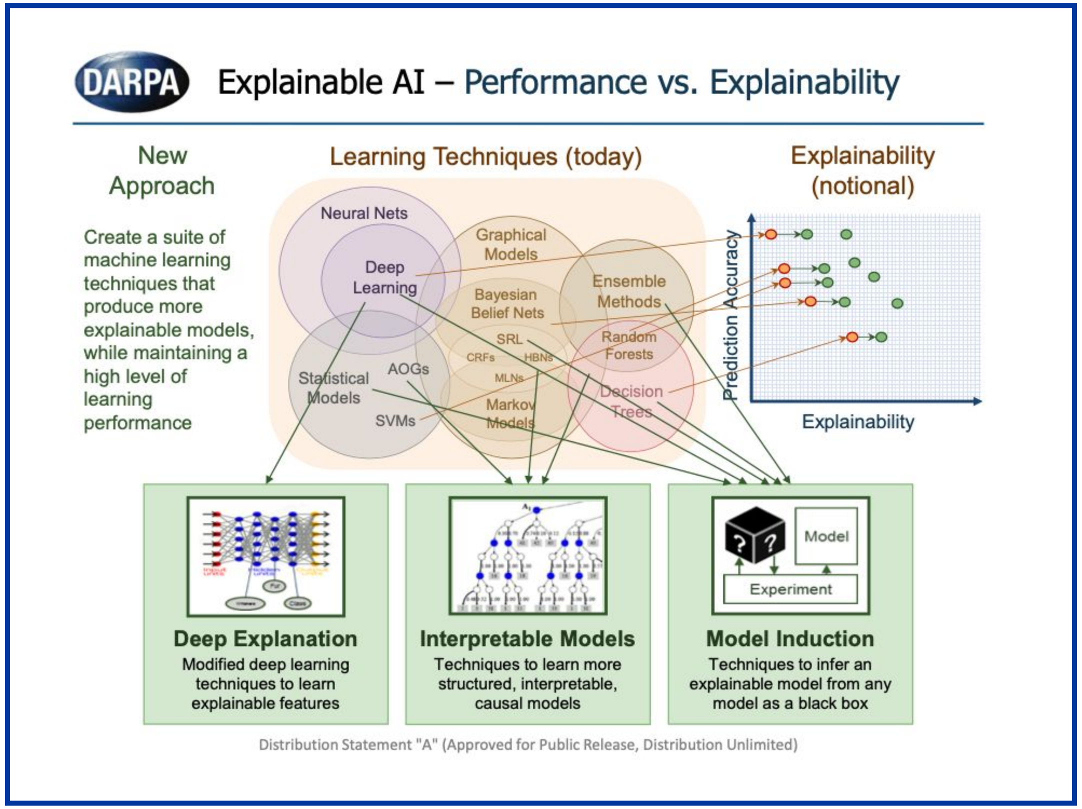
VECTOR INSTITUTE | UNIVERSITY of GUELPH

https://www.theverge.com/2018/1/12/16882408/google-racist-gorilla s-photo-recognition-algorithm-ai

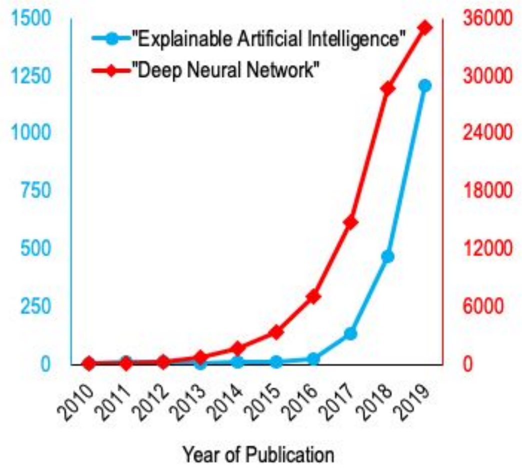
Table of Contents

A Review of XAI

Recent Origins & Progress



- XAI goes back before deep learning, but the recent surge is traceable to DARPA's initiative
- Gunning's presentation was published and picked up by highly visible popular press stories



Gunning, D. (2017). Explainable Artificial Intelligence (XAI)—DARPA. *Machine Learning*, 18.

A Review of XAI

An Early Taxonomy

1
- Proxy Models

2
- Introspective Models

3
- Correlative Techniques & Saliency Maps

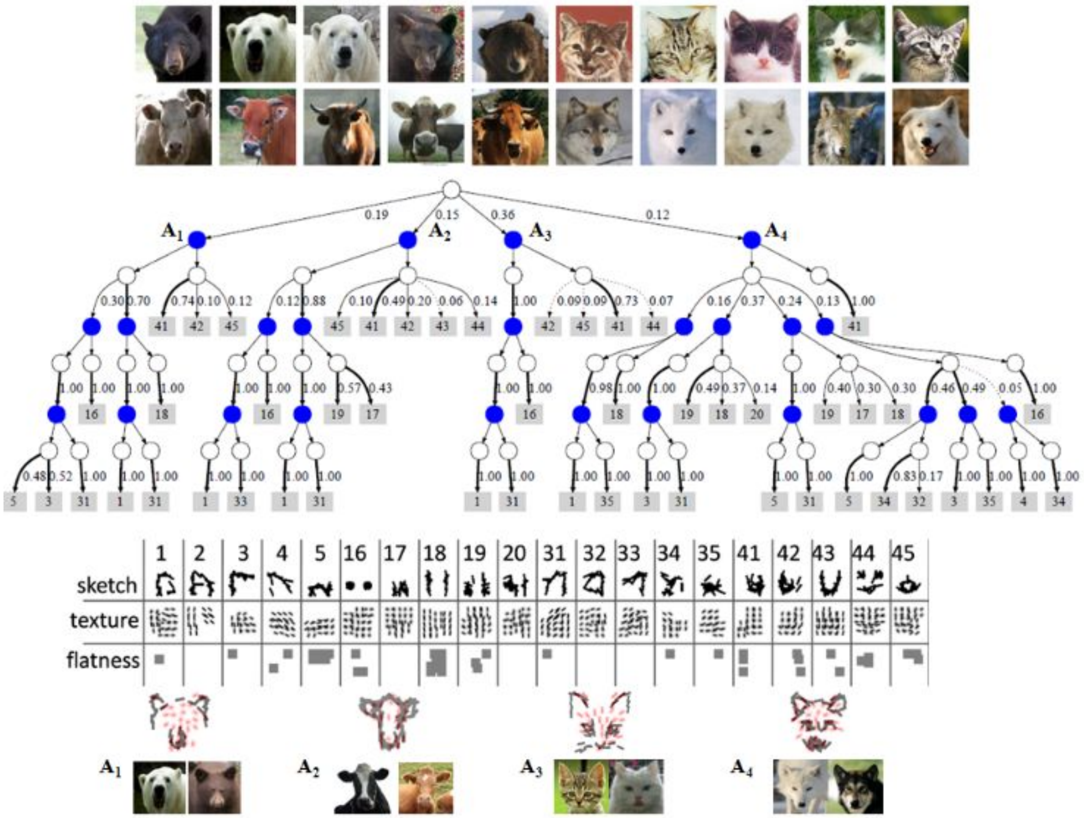
4
- Post Hoc Explanations

5
- Example-Based Explanations

VECTOR INSTITUTE | UNIVERSITY of GUELPH

A Review of XAI

Proxy Models



Si, Z., & Zhu, S.-C. (2013). Learning AND-OR Templates for Object Recognition and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(9), 2189–2205.

- Proxy models train an adjacent model with a simpler, interpretable architecture to express what the more complicated NN is doing
- In this case, the proxy model learns a set of and / or rules to apply to simple features of input images, ultimately resulting in a logic tree that expresses how the NN chooses its class

Pros:

- Proxy models use highly interpretable architecture (e.g. decision trees)

Cons:

- Proxy models cannot match the performance of the model they are standing in for

A Review of XAI

Introspective Models

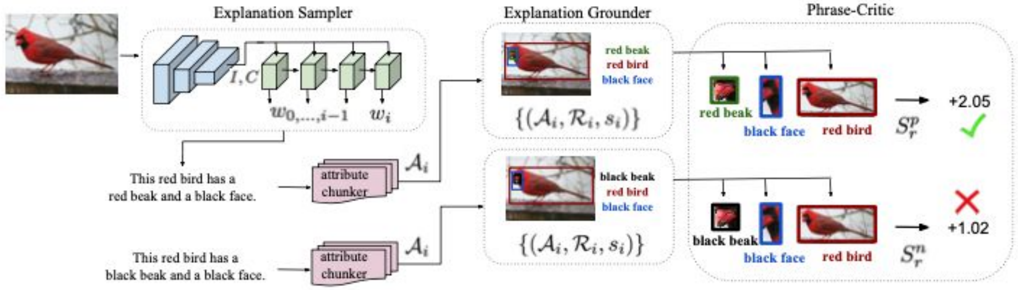
- Introspective models append a secondary DNN to the one being explained to learn to express its decisions in interpretable output (e.g. language)

Pros:

- Introspective models present the user with compelling explanations that imply causation
- No loss in predictive power

Cons:

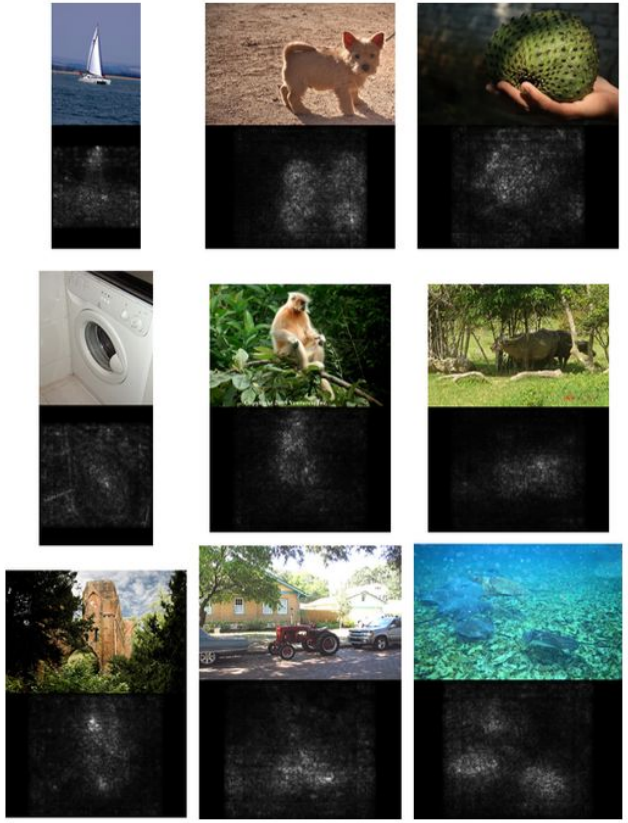
- Replaces one black box with another; who explains the explainer?



Hendricks, L. A., Hu, R., Darrell, T., & Akata, Z. (2018). Grounding Visual Explanations. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11206, pp. 269–286).

A Review of XAI

Correlative Methods & Saliency Maps



- These techniques visualize some relationship between the input and output without changing the model
- Often visualize the loss gradient with respect to the input

Pros:

- Directly describes relationship between input and output
- Intuitive format

Cons:

- Gradients are correlated with optimal loss, but they can point to local minima
- Point to a manifold of explanations with no guidance on narrowing the field
- Susceptible to misinterpretation – indistinguishable from edge detectors in some cases

VECTOR INSTITUTE | UNIVERSITY of GUELPH

(1) Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps.
(2) Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (n.d.). *Sanity Checks for Saliency Maps*. 11.

A Review of XAI

Post Hoc Explanations

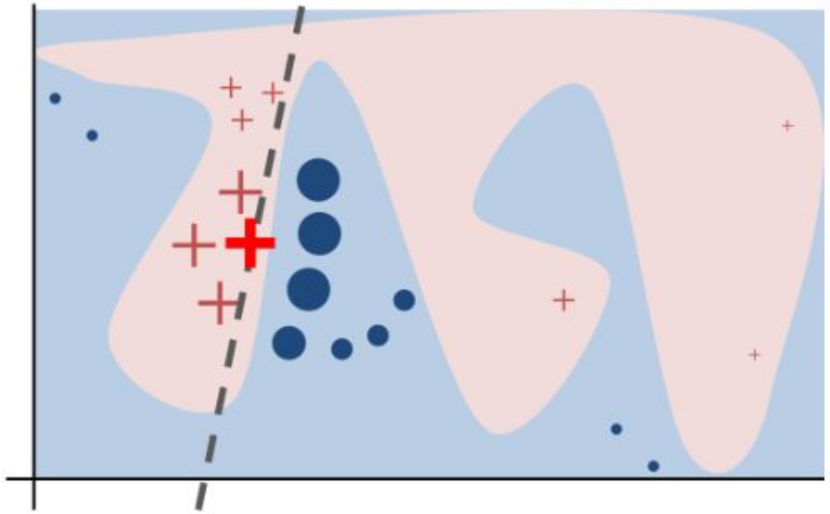
- These include any technique that asks the user to rationalize the cause of the explanation after having observed it
- In LIME, the explanations are local; any pixel not weighted heavily in the linear approximation around a certain point are omitted from the explanation

<u>Pros</u>:

- Can be causal, based on iterative perturbation or processing bottlenecks (e.g. attention)

<u>Cons</u>:

- Susceptible to the confirmation bias in interpretation
- Susceptible to bias in defining regions of interest, hyperparameters of explanation



VECTOR INSTITUTE | UNIVERSITY of GUELPH

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier.

A Review of XAI

Example-Based Explanations

- These methods offer explanations in the form of examples that illustrate when the model behaves certain ways
- E.g. here is an input case that is highly representative of this model's view of its class / here is an input case that is an exception

Pros:

- Involves abductive logic; shows when models' failure in addition to success and sketches the decision boundary through cases

Cons:

- Examples are generated from a massive stimulus space, so explanations are generated by users ad hoc
- No guidance on narrowing down explanations, involves many researcher degrees of freedom

VECTOR INSTITUTE | UNIVERSITY of GUELPH

Kim, B. (2016). Examples are not enough, learn to criticize! Criticism for Interpretability.

A Review of XAI

What's Missing?

The categories reviewed above are powerful techniques that offer a lot of explanatory power

However, the current state of XAI has a collective blind spot:

-> Most explanations are generated *post hoc* rather than *a priori*

-> Most investigations are confirmatory rather than falsifiable

-> Most explanations are automatic, with many researcher degrees of freedom

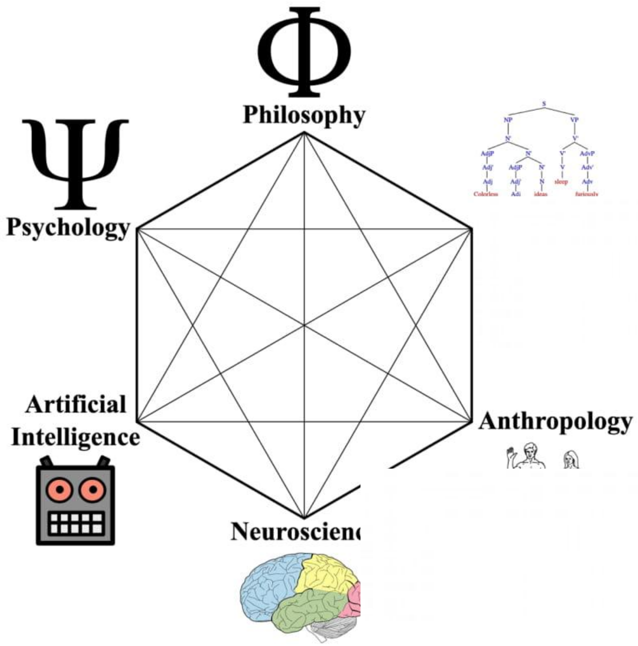
A complementary approach is to conduct experiments that test hypotheses under falsifying conditions with curated controls

Table of Contents

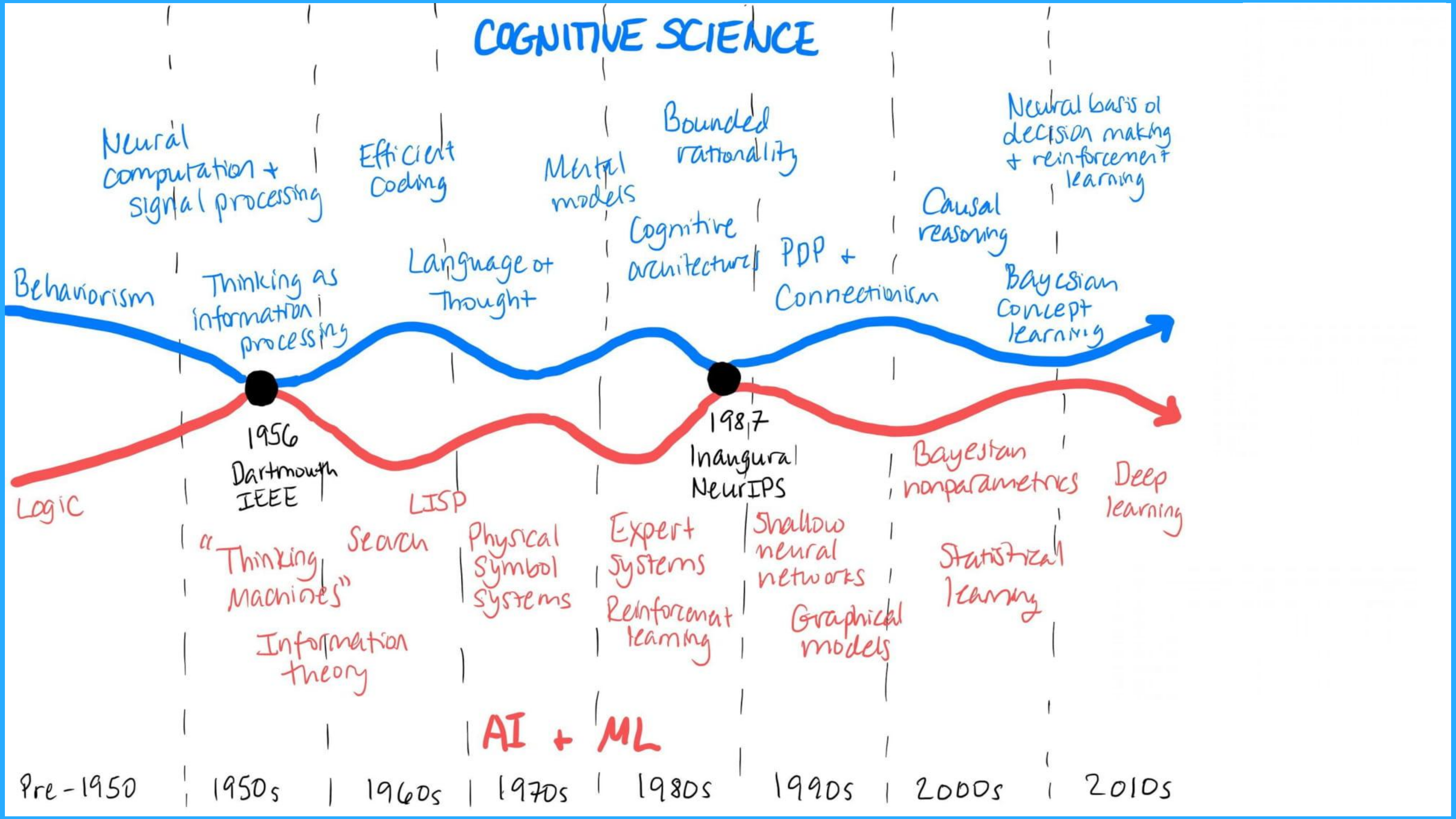
Cognitive Science

What is Cognitive Science?

*The **study of intelligent systems** and how they produce behavior, rooted in the assumption that those systems follow **principles of computation**.*



Cognitive Science and AI



Artificial Cognition

A Branch of Machine Behaviour towards XAI

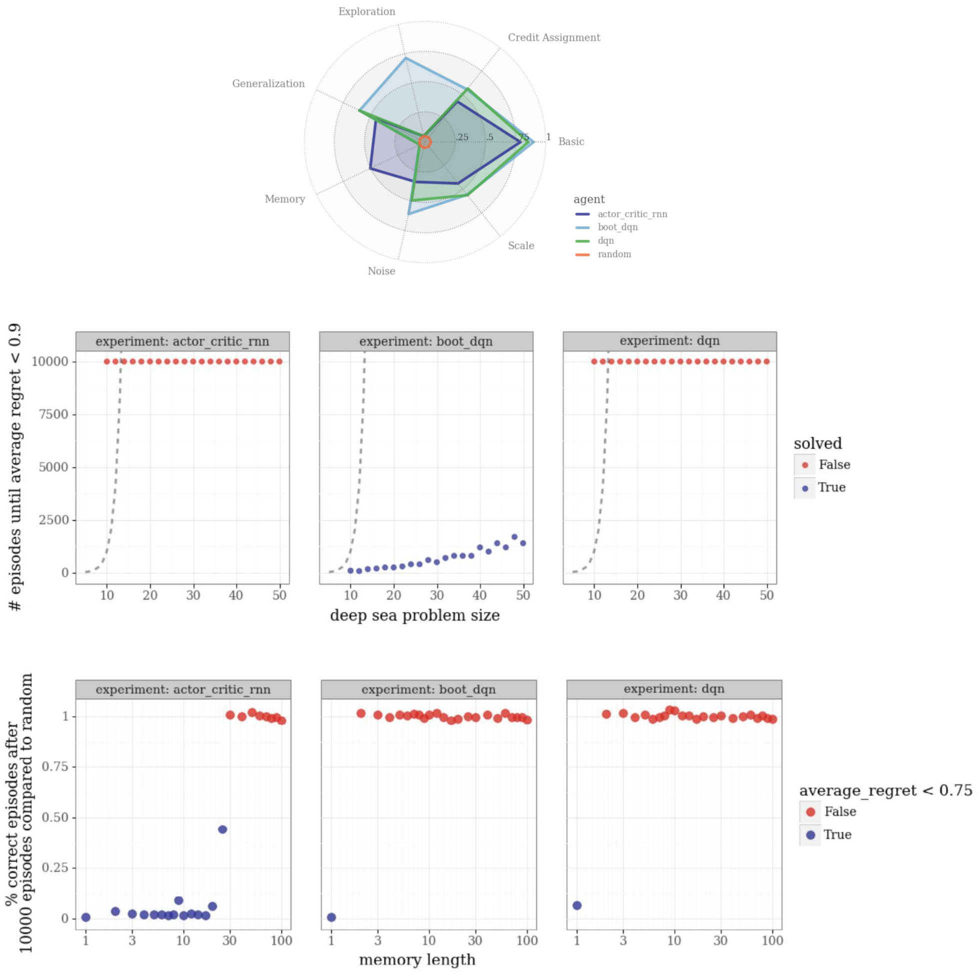
Framework for Study:

1. Document Variations in Behaviour
2. Infer the Cause by Falsifying Alternatives
3. Identify Boundary Conditions
4. Toy with the Brain

S. Ritter, D. G. T. Barrett, A. Santoro, and M. M. Botvinick. Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study. page 10, 2017.

Artificial Cognition

1. Document Variations in Behaviour



Osband, I., Doron, Y., Hessel, M., Aslanides, J., Sezener, E., Saraiva, A., McKinney, K., Lattimore, T., Szepezvari, C., Singh, S., Van Roy, B., Sutton, R., Silver, D., & Van Hasselt, H. (2019). Behaviour Suite for Reinforcement Learning.

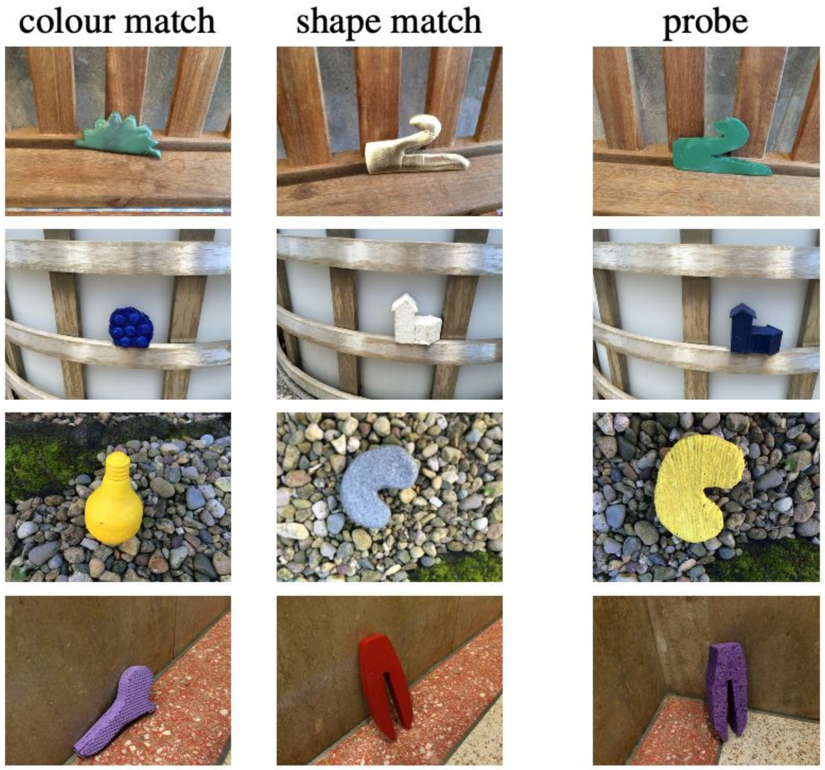
- To develop testable hypotheses, we don't want to take shots in the dark
- Start from documented variations; correlate behaviours with tasks
- Osband et al (2019) created a set of 7 RL benchmarking tasks that explicitly load onto different behaviours and tested them on 3 different agents with different architectures
- This allowed them to test specific hypotheses about how different architectures would perform on different tasks (e.g. DQN better explorations vs A2C RNN better memory)

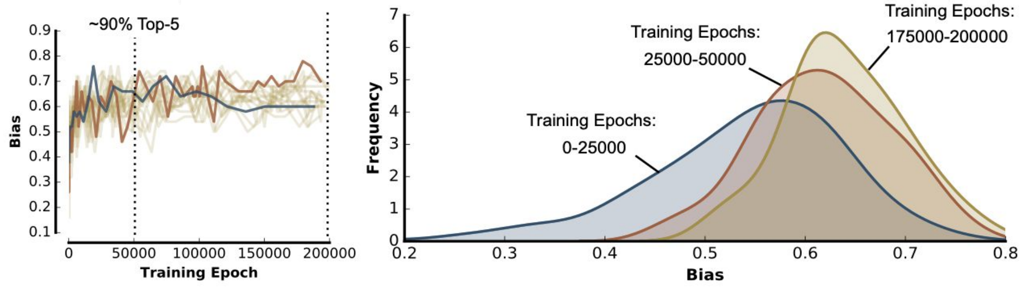
VECTOR INSTITUTE | UNIVERSITY of GUELPH

Artificial Cognition

2. Infer the Cause

- Wanted to know whether ANNs exhibit shape bias, which is the human tendency to over-index on shape versus colour when learning new objects
- Used a test set from human development psychology used to assess which pairs learners find more similar (control for background etc)
- Used a nearest-neighbour algorithm to measure the network’s preference for shape or colour
- Strong preference for shape-matching probes over colour-matching controls



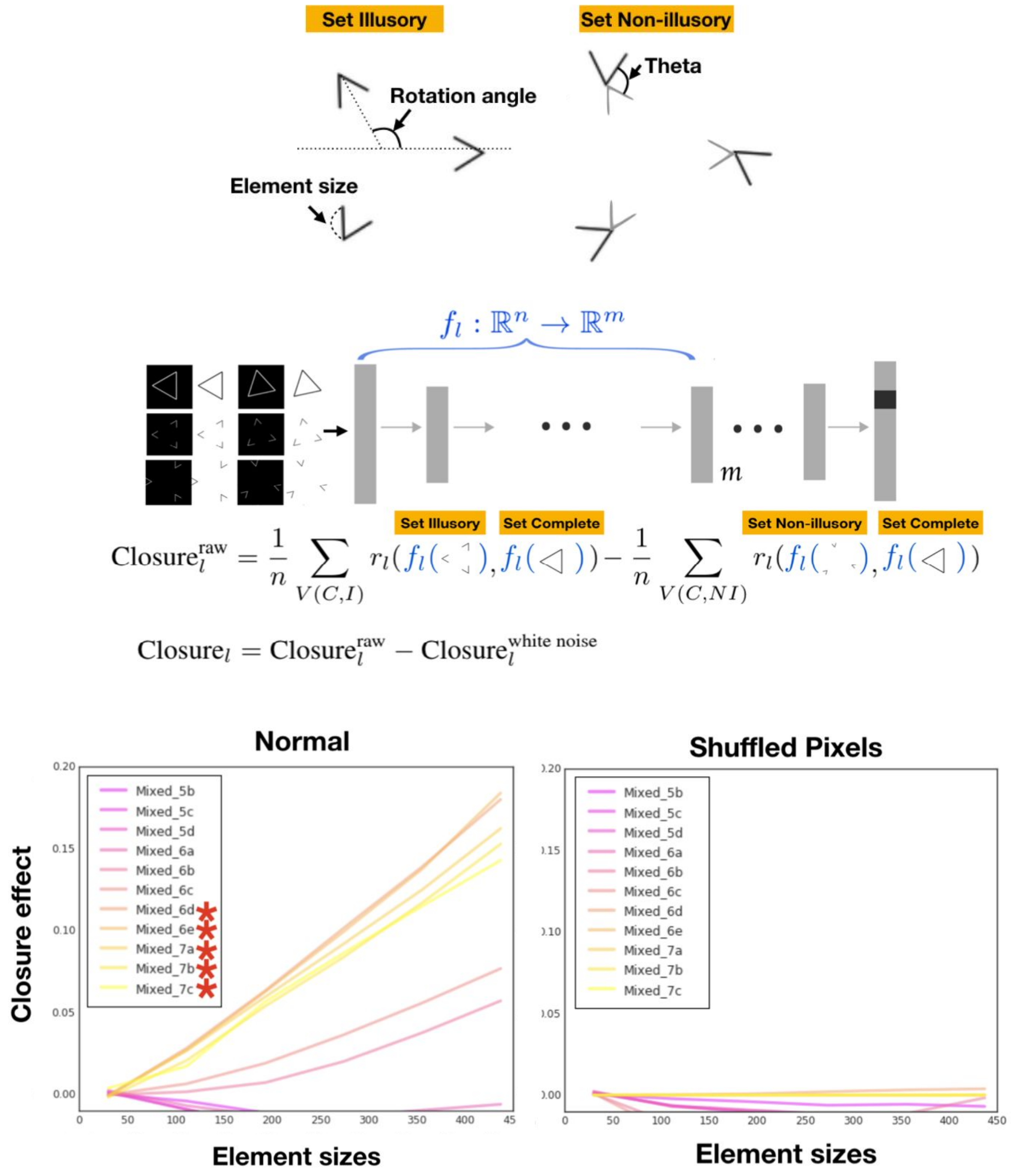


Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). *Cognitive Psychology fo Deep Neural Networks: A Shape Bias Case Study*. 10.

Artificial Cognition

2. Infer the Cause

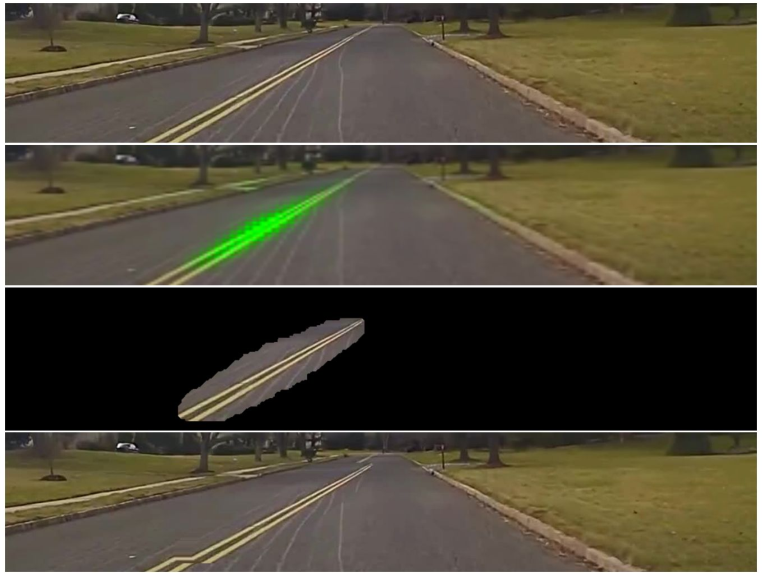
- Wanted to know whether ANNs exhibit Gestalt closure, which is a behaviour in biological NNs to view incomplete shapes as whole
- Developed a closure metric, which compares the cosine similarity of internal layers' output between full triangles and illusory or non-illusory (rotated vertices) triangles
- Tested 7 different hypotheses
- Here, they fail to reject the hypothesis that later layers exhibit stronger closure than earlier layers
- Used shuffled pixels and three other controls

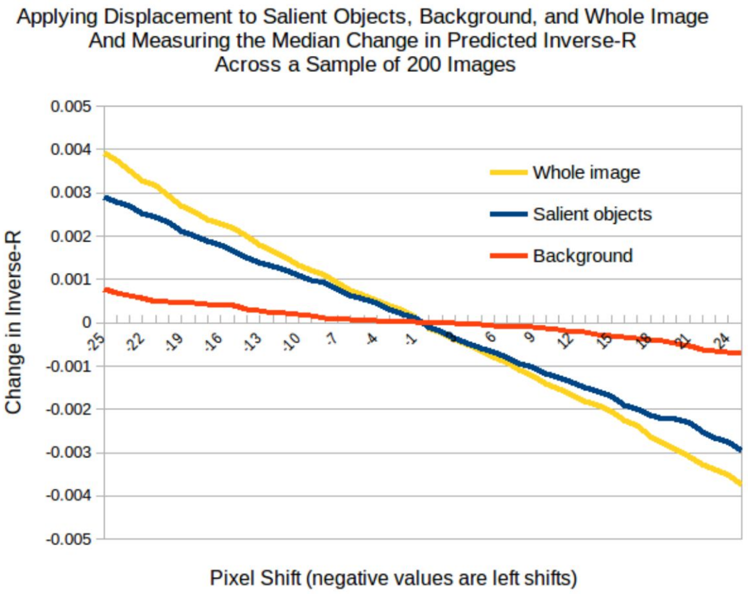


Artificial Cognition

2. Infer the Cause

- Developed a saliency algorithm to highlight the visual input to their steering algorithm that ought to correspond to steering output
- Recognizing that there is correlation with ground contours, the authors wanted to test whether the highlighted portions affect the steering angle
- Created a set of input stimuli with displaced pixels (salient/background/entire image) to rule out alternatives
- Show that displacing the critical pixels is equivalent to displacing the entire image, but only in the presence of a background



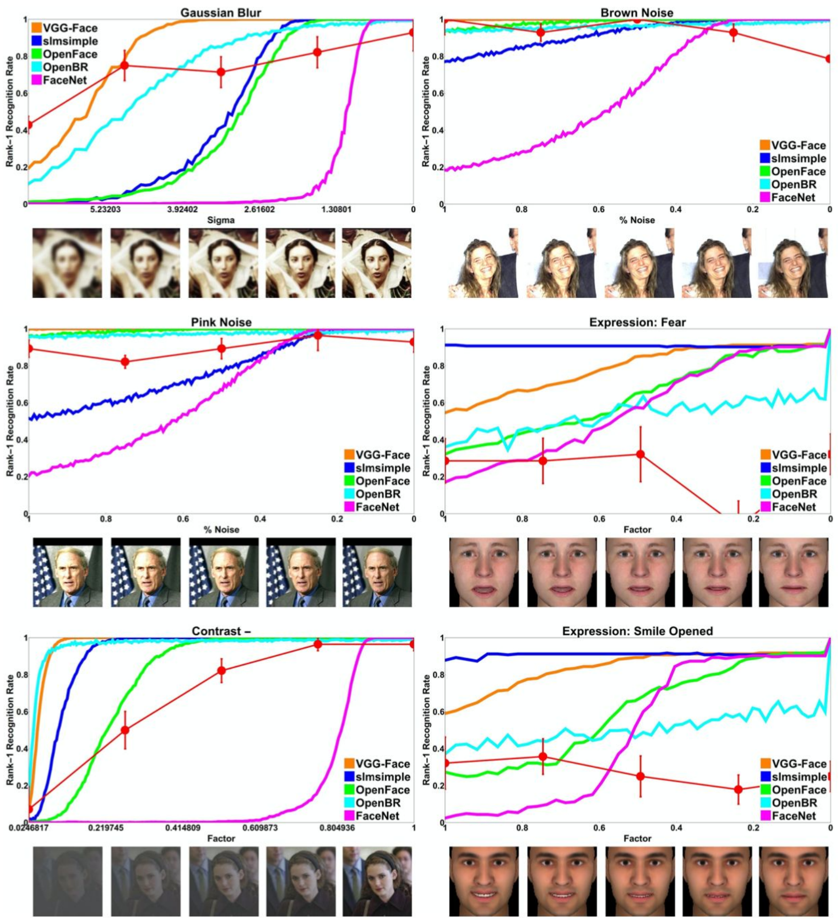


VECTOR INSTITUTE | UNIVERSITY of GUELPH

Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining How a Deep Neural Network Trained with End-to-End Learning Steers a Car.

Artificial Cognition

3. Identify Boundary Conditions



- If your theory can explain when a behaviour happens, it should also account for when it stops; important to narrow the range of viable alternative explanations
- RichardWebster et al. (2018) applied a set of perturbations across a range of intensities to 5 different face recognition models (including expression)
- One of the neat findings from this explorative study was that FaceNet and OpenFace, which are variants of the same architecture, performed very differently
  - "FaceNet uses a subset of MS-Celeb-1M where difficult images that contain partial occlusion, silhouettes, etc. have been removed as a function of facial land- mark detection. This is likely the weakest link, as the network does not have an opportunity to learn invariance to these conditions."

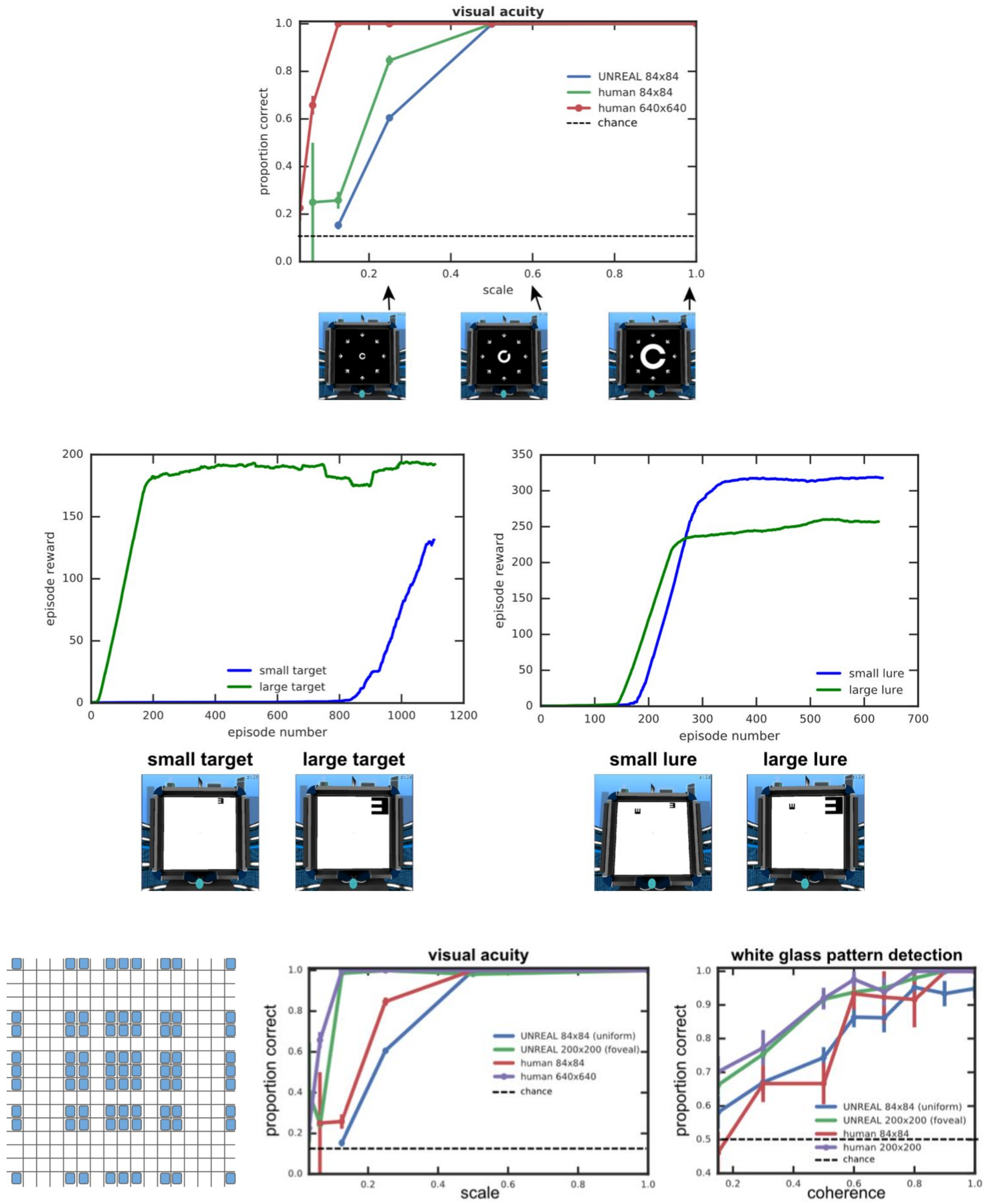
VECTOR INSTITUTE | UNIVERSITY of GUELPH

RichardWebster, B., Kwon, S. Y., Clarizio, C., Anthony, S. E., & Scheirer, W. J. (2018). Visual Psychophysics for Making Face Recognition Algorithms More Explainable. In V. Ferrari, M. Hebert, C. Sminchisescu, & Y. Weiss (Eds.), *Computer Vision – ECCV 2018* (Vol. 11219, pp. 263–281).

Artificial Cognition

4. Toy with the Brain

- Normally not possible with humans, ML researchers can learn from experimentation by altering the "brain"
- Leibo et al. (2018) put UNREAL RL agent in a virtual environment populated by experimental stimuli from visual psychophysics
- Showed exemplary performance on most things except visual acuity (clarity of detail) and contrast
- They then predicted that UNREAL would have a had time learning small relative to large items, and would be disproportionately distracted by large items
- Corrected this flaw after designing a new input filter inspired by the human fovea



VECTOR INSTITUTE | UNIVERSITY of GUELPH

Leibo, J. Z., d'Autume, C. de M., Zoran, D., Amos, D., Beattie, C., Anderson, K., Castañeda, A. G., Sanchez, M., Green, S., Gruslys, A., Legg, S., Hassabis, D., & Botvinick, M. M. (2018). Psychlab: A Psychology Laboratory for Deep Reinforcement Learning Agents.

Table of Contents

VECTOR INSTITUTE | UNIVERSITY of GUELPH

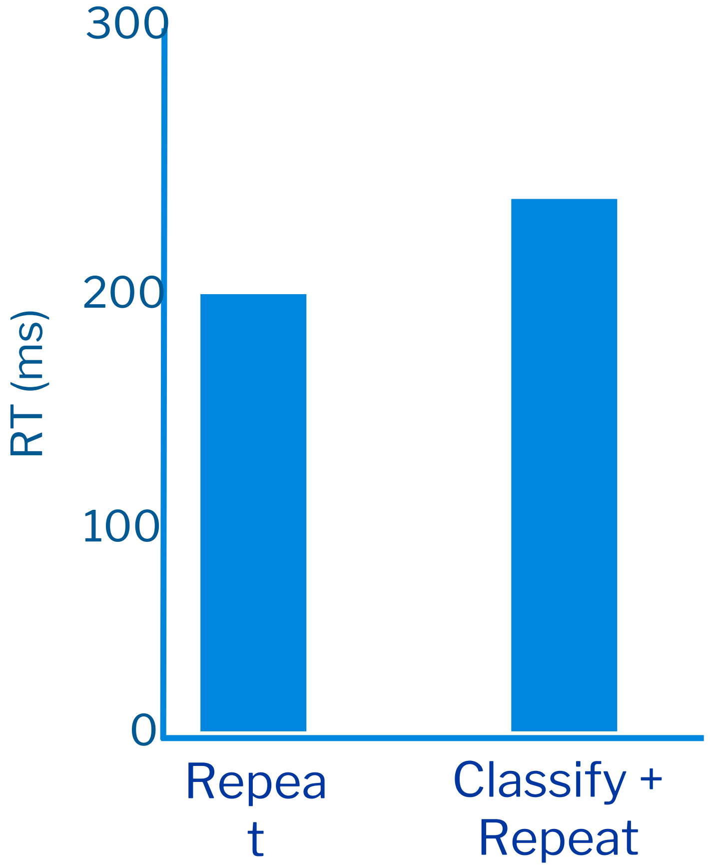
Response Time Methods for XAI

How would you explain AI if you couldn't look inside the black box?

- Dominant XAI techniques require some way to query the model's architecture, parameters, gradient, etc.
- For many important XAI cases, researchers will not have privileged access to the model in question
- We want a proof of concept for an explanation derived strictly from a priori hypotheses about the output given the input; no peeking inside!
- The challenge is that the output (label, accuracy) does not have an obvious relationship to the internal processes

RT Methods for XAI

Explaining Human Vision

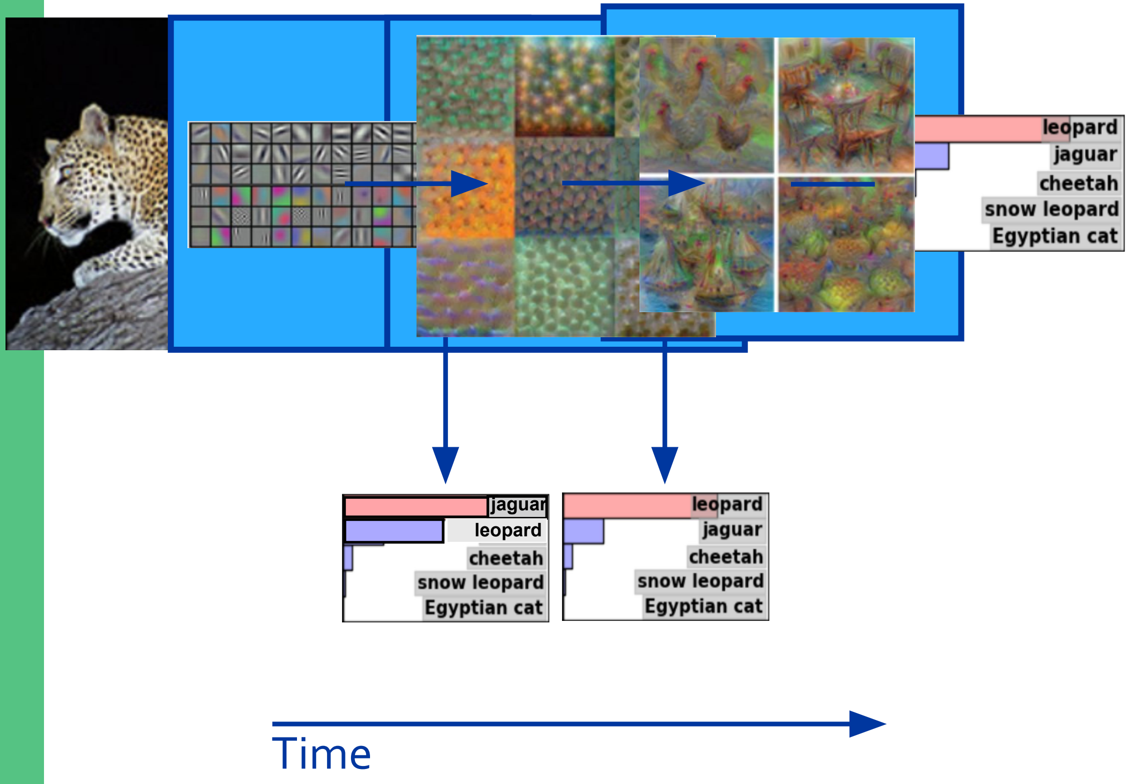


- Psychologists also have a black-box problem in explainability
- Before modern neuroimaging, psychologists were unable to look inside their black box
- RT methods were invented in 1868 to identify different stages of perceptual processing
- By carefully manipulating the input or task, experimenters could attribute differences in RT to otherwise hidden processes
- To use RT methods, we require a distribution of RTs and a meaningful connection between processing time and performance

Donders, F. C. (tr. 1969; or. 1868). On the speed of mental processes. *Acta psychologica*, *30*, 412-431.

RT Methods for XAI

Dynamic Inference

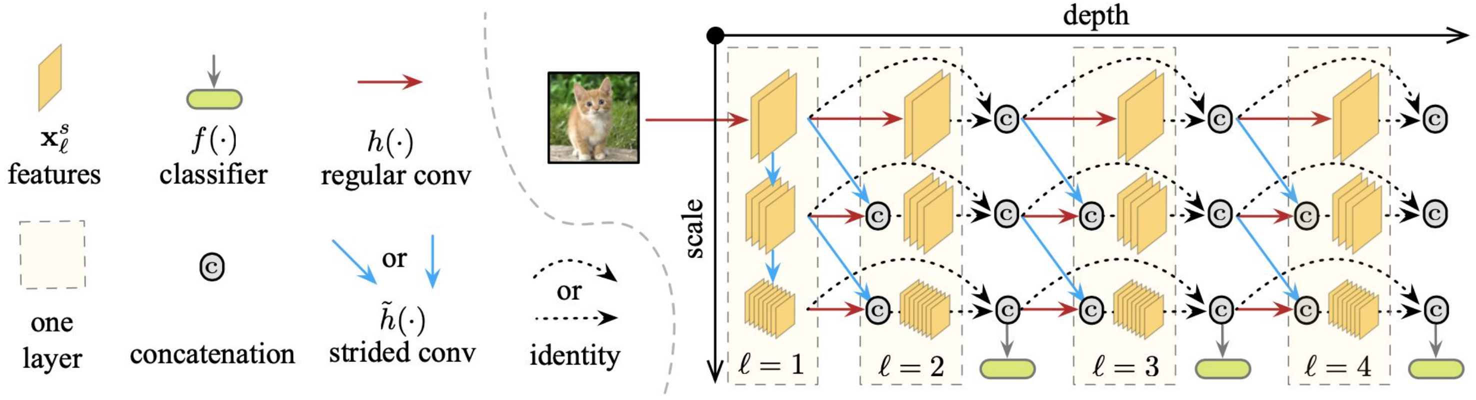


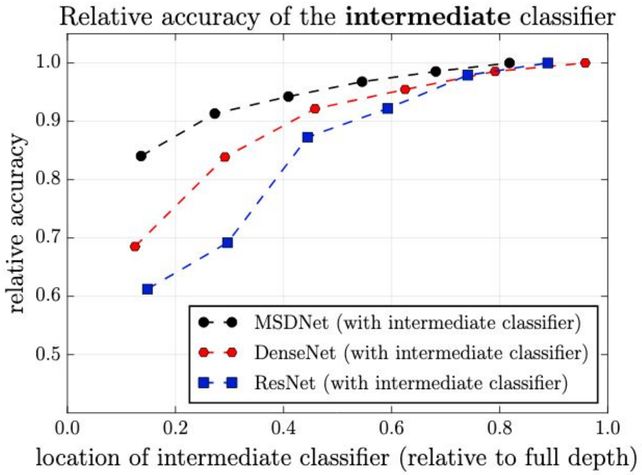
- One solution is dynamic inference models, which permit early-exits based on the confidence of intermediate classifiers
- These models are gaining popularity as the demand grows for devices with:
  - Limited computational capacity
  - Time-constrained decision-making
- This produces two conditions required for RT methods:
  - Variability in RT
  - Meaningful connection between RT and performance

VECTOR INSTITUTE | UNIVERSITY of GUELPH

RT Methods for XAI

Dynamic Inference





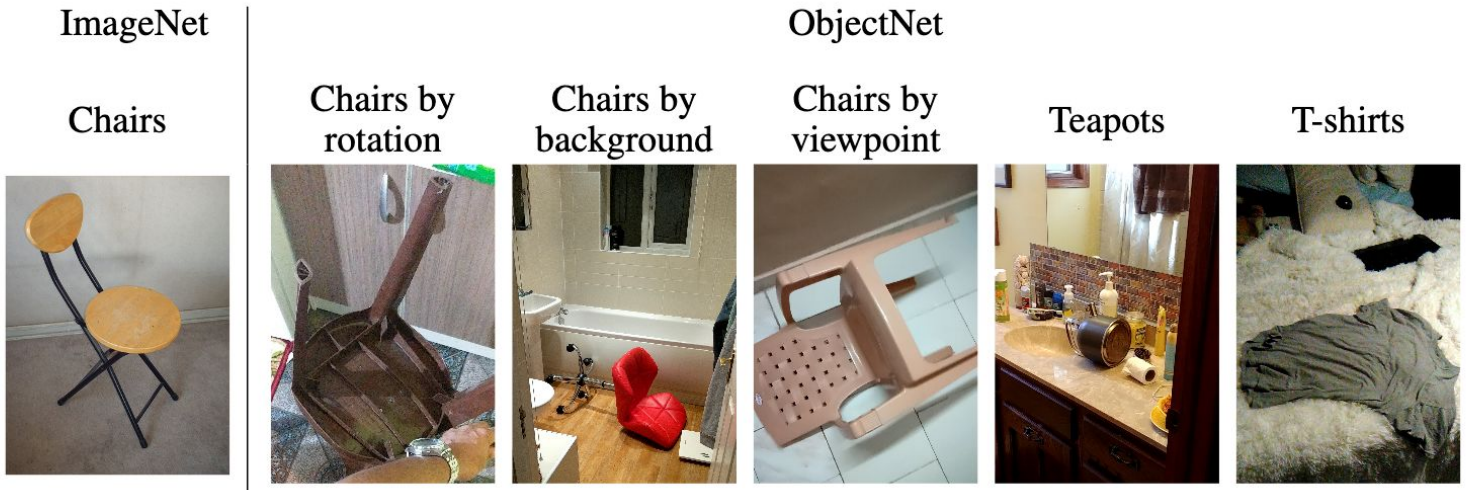
- If hierarchical feature space is correlated with model depth, and conditional computation allows early exits, then we can make predictions about feature space and RT
- Not a perfect correlation because the architecture does permit sharing features between layers
- Specifically, decisions that depend on higher-order feature space should take longer
- RT is handy because it is completely "outside" the black box

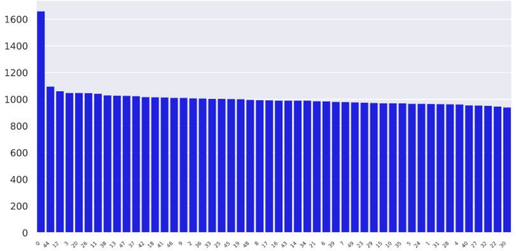
VECTOR INSTITUTE | UNIVERSITY of GUELPH

Huang, G., Chen, D., Li, T., Wu, F., Van Der Maaten, L., & Weinberger, K. Q. (2017). Multi-scale dense convolutional networks for efficient prediction. *arXiv preprint arXiv:1703.09844*, 2.

RT Methods for XAI

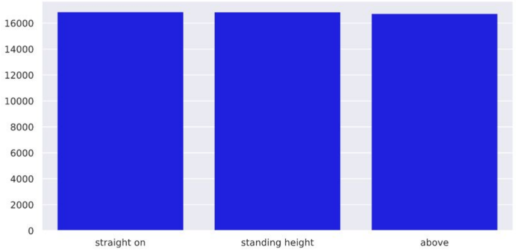
Experiment 1 - Method





Rotation*

Background



Viewpoint

- ImageNet overrepresents canonical image features
- ObjectNet deliberately includes complex and unusual features

- A 50 000 – image test set for object recognition algorithms that contains non-canonical viewpoints, backgrounds, and full rotation

VECTOR INSTITUTE | UNIVERSITY of GUELPH

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., ... & Katz, B. (2019). ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems* (pp. 9448-9458).

RT Methods for XAI

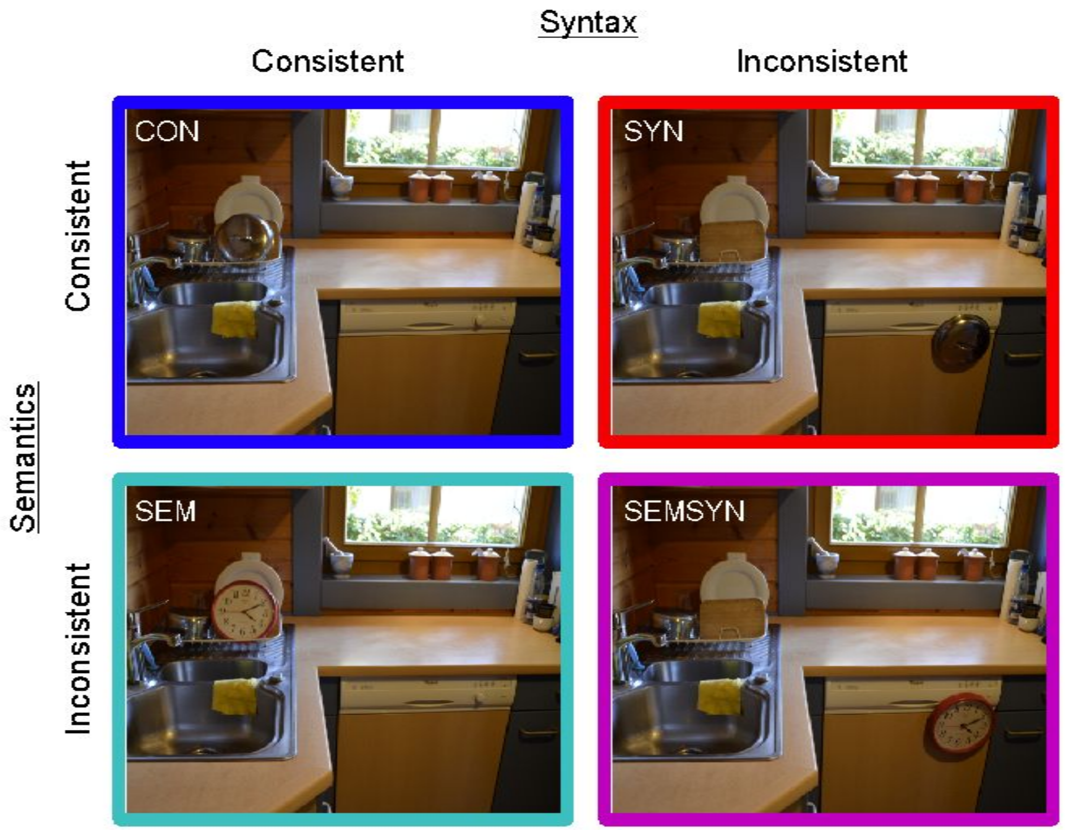
Experiment 1 - Results

VECTOR INSTITUTE | UNIVERSITY of GUELPH

RT Methods for XAI

Experiment 2 - Method

- SCEGRAM database is a 62-image test set designed for experiments with humans
- Carefully controls saliency, position, and size of the critical object while varying the scene's semantics or syntax



Öhlschläger, S., & Võ, M. L. H. (2017). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior research methods*, *49*(5), 1780-1791.

RT Methods for XAI

SCEGRAM - Consistent

VECTOR INSTITUTE | UNIVERSITY of GUELPH

RT Methods for XAI

SCEGRAM – Semantic Inconsistency

RT Methods for XAI

SCEGRAM – Syntactic Inconsistency

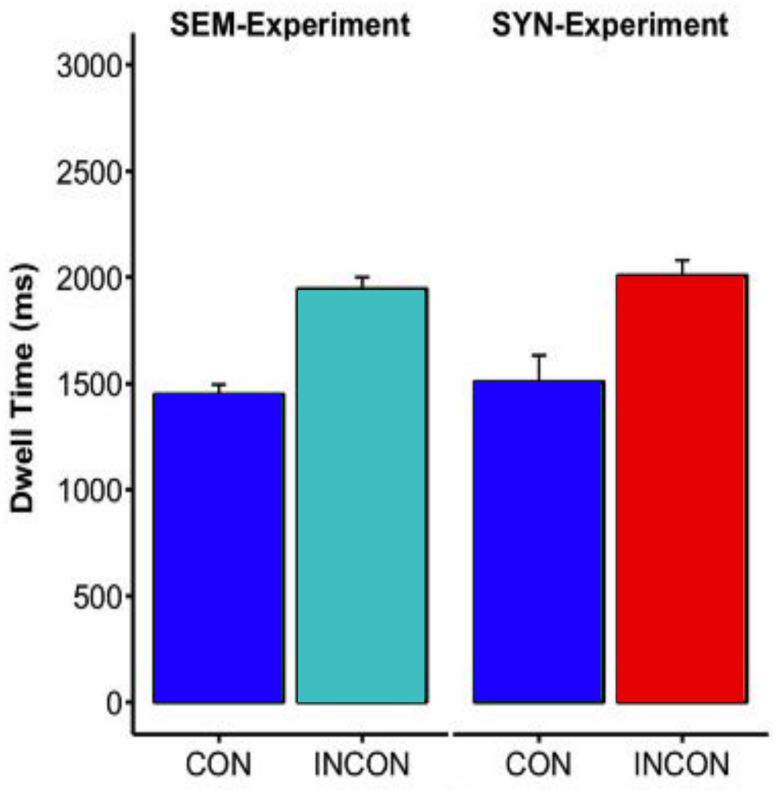
RT Methods for XAI

SCEGRAM – Semantic & Syntactic Inconsistency

RT Methods for XAI

Experiment 2 - Background

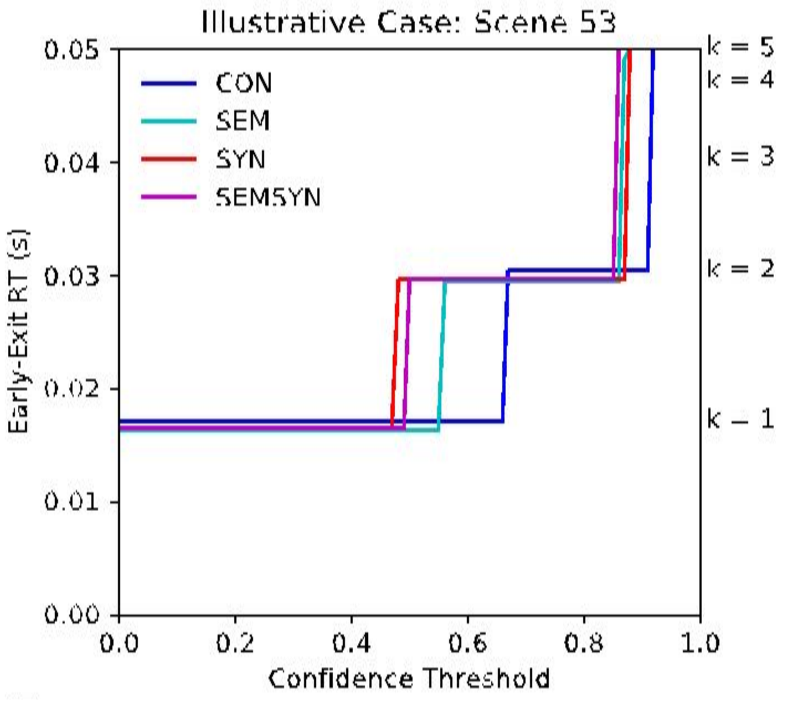
- Scene grammar is a human phenomenon whereby the visual system is very sensitive to high-level semantic and syntactic relationships between objects and the scene they appear in
- We have an easier time processing scenes with consistent grammar
- Attention is attracted to violations and spends more time processing them

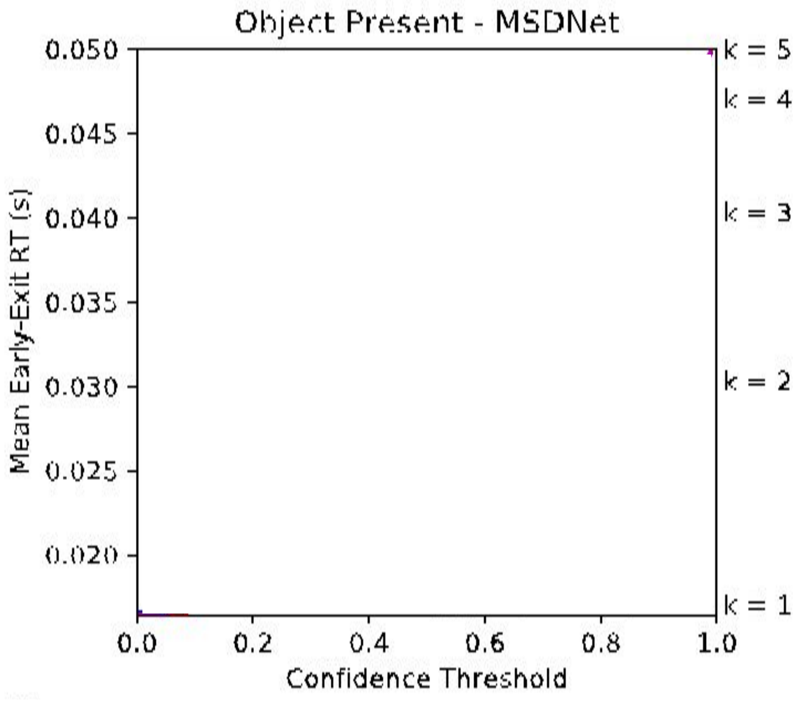


Öhlschläger, S., & Võ, M. L. H. (2017). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior research methods*, *49*(5), 1780-1791.

RT Methods for XAI

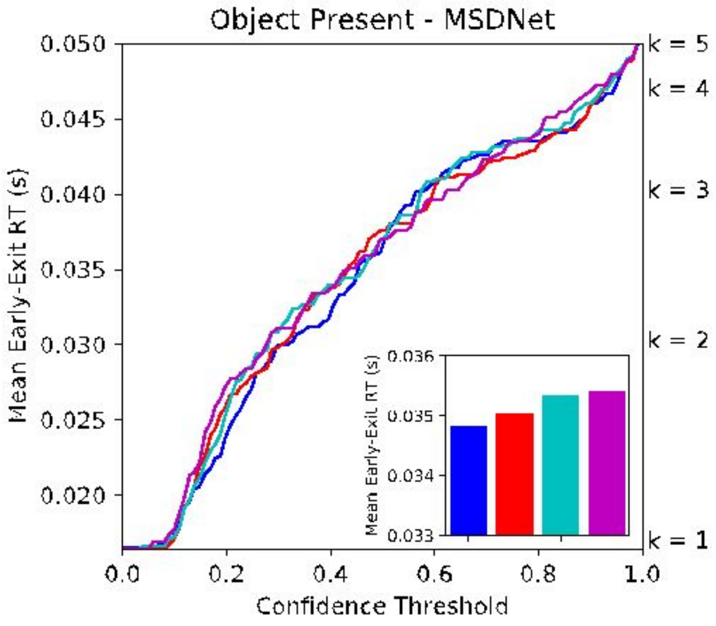
Experiment 2 - Results

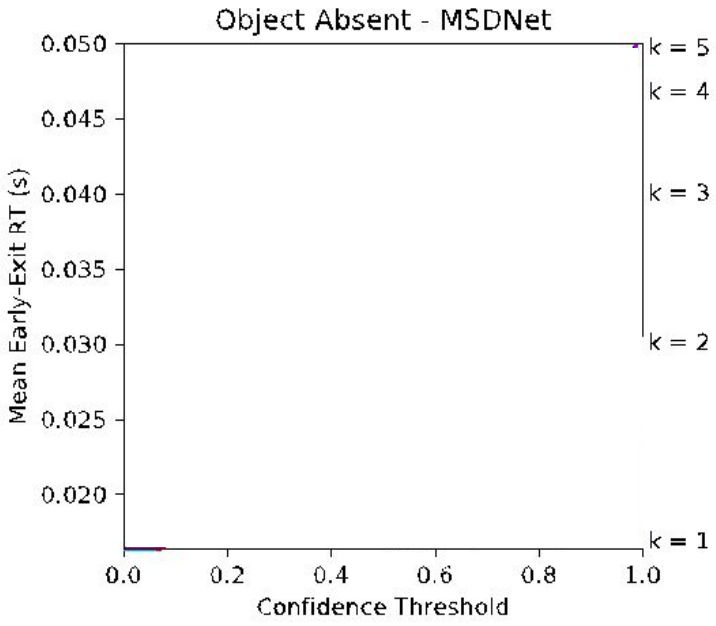




RT Methods for XAI

Experiment 2 - Results





- Object-Absent images are, by definition, semantically consistent – they match CON
- As predicted, the lack of object-scene violations is reflected in homogenous RT

Conclusions

- Response time analyses can be used to make inferences about CNN feature processing in dynamic inference models from completely "outside" the black box
- These techniques lend themselves to *a priori* hypothesis testing about the relationship between the input space and model behaviour
- These analyses could be used to form expectations for when and how models should perform in situations where explanations are desirable, but privileged access to a model is denied.

Resources

- AI + Cognitive Science: https://cbmm.mit.edu/learning-hub
- Interpretability:
  - Book https://christophm.github.io/interpretable-ml-book/
  - Blog posts, visualizations and code https://distill.pub/
  -