Whitebox AI

ML meets Animal Communication

Hunaid Hameed

Interpretability

- Miller (2017): "Interpretability is the degree to which a human can understand the cause of a decision."
- Why do we need interpretability:
    - Curiosity
    - Audit and improvement
- Interpretability can be model-specific and model agnostic.

Scope

- Algorithm Transparency: Knowledge of the algorithm not data or model e.g. Logistic Regression
- Holistic Interpretability: Understand it all *holistically* e.g. Naive Bayes Probability
- Modular Understanding: Understanding parts of the whole e.g. parts of decision tree
- Interpretability for a Single Prediction: Infer decision for single a instance
- Interpretability for a Group of Prediction: Infer decision for a group of instances

But interpretability is not explainability...

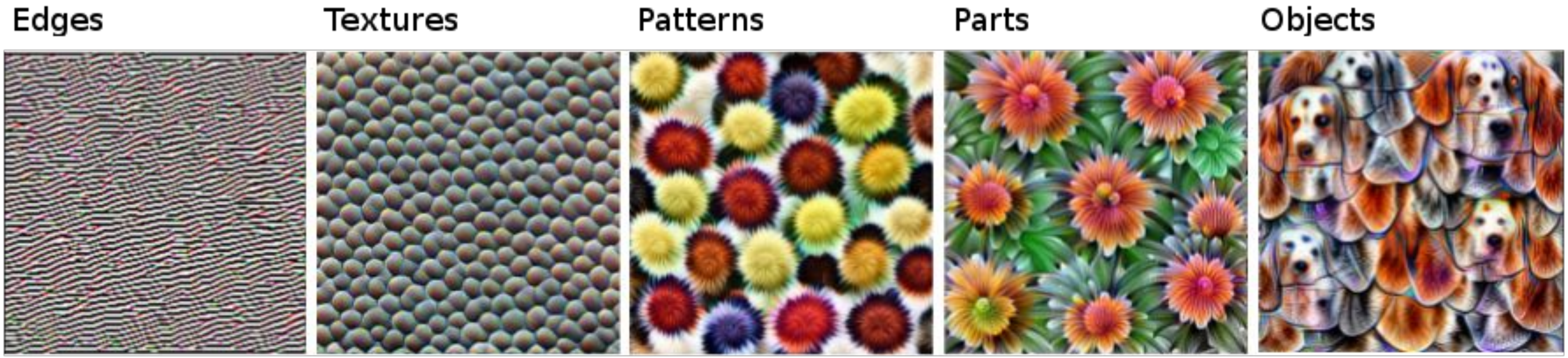
It does not answer ‘why’

We can answer the question of *what*, may be even *how*, but now *why*

We cannot contrast and answer questions of: *what not* and *how not*

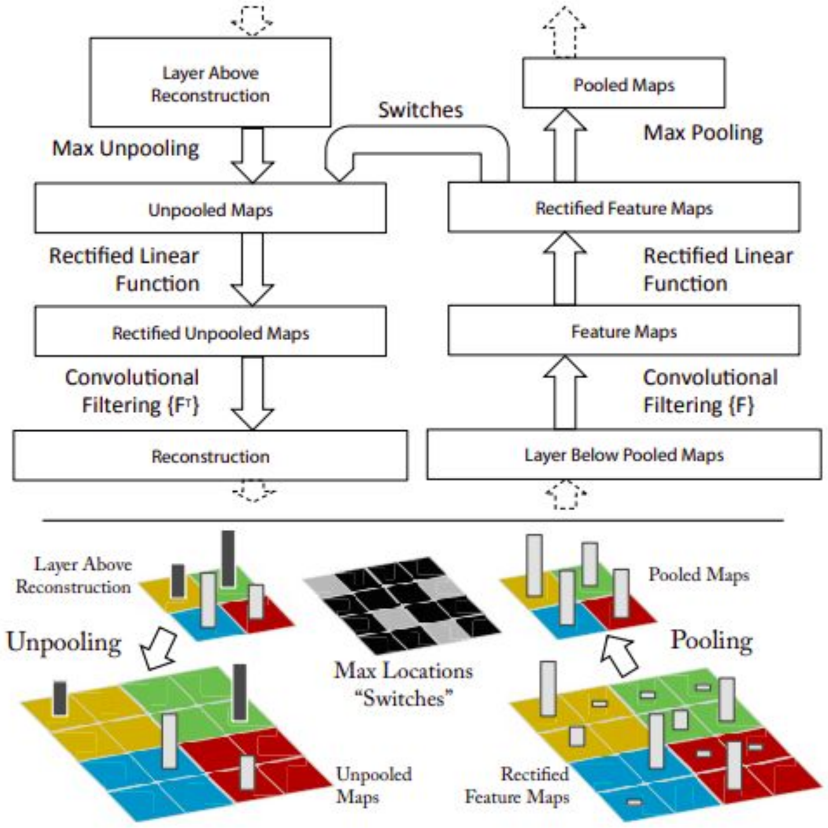
Interpretability in Neural Networks

- Lower layers learn basic features and shapes
- As we move further the shapes become more complex
- This is also seen in the visual pathways of the human brain



Deconvolution

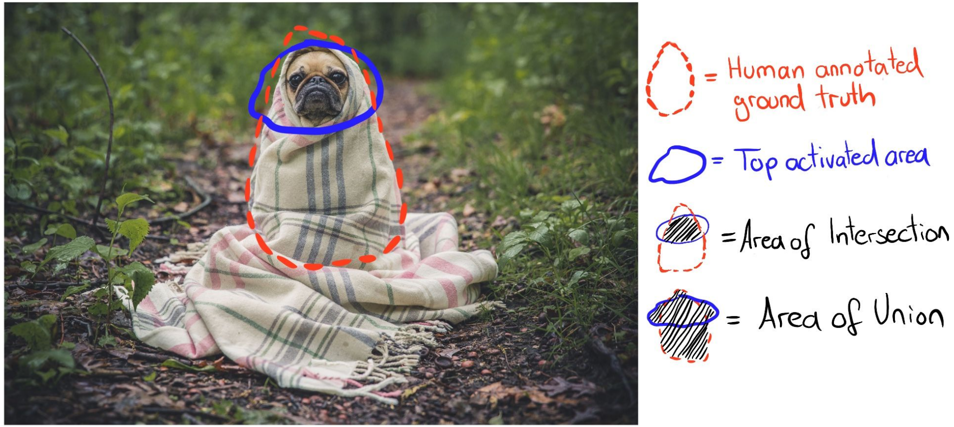
- Work of Matthew Zeiler and Fergus at NYU
- Method:
    - Feedforward input into a DNN
    - Identify neuron which is highest activated
    - Turn off (make 0) all other activations from layer of neuron
    - Feedback this matrix
    - Visualize output

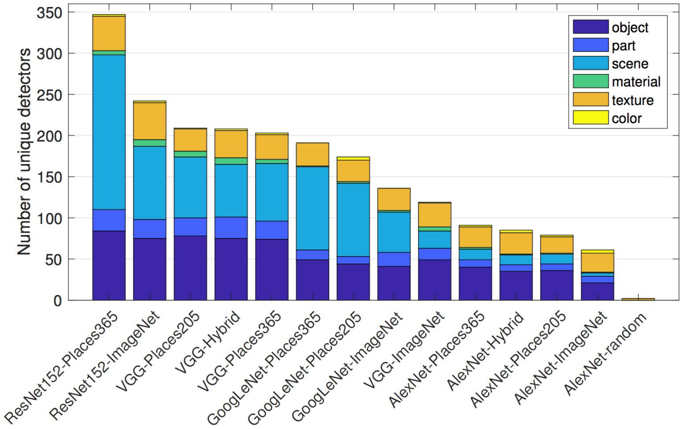


Max and Min Deconvolution

Network Dissection

- By Bau & Zou et al. (2017)
- Algorithm:
  - Obtain neuron activation for an image
  - Find the top activation for a part of a layer
  - Map it to the actual image
  - Determine Intersection over Union (IoU)

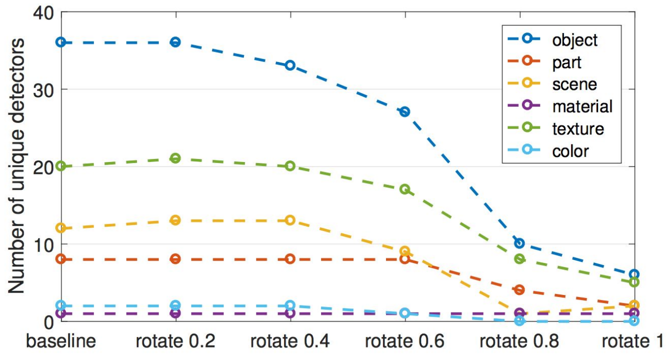




Concept Detector

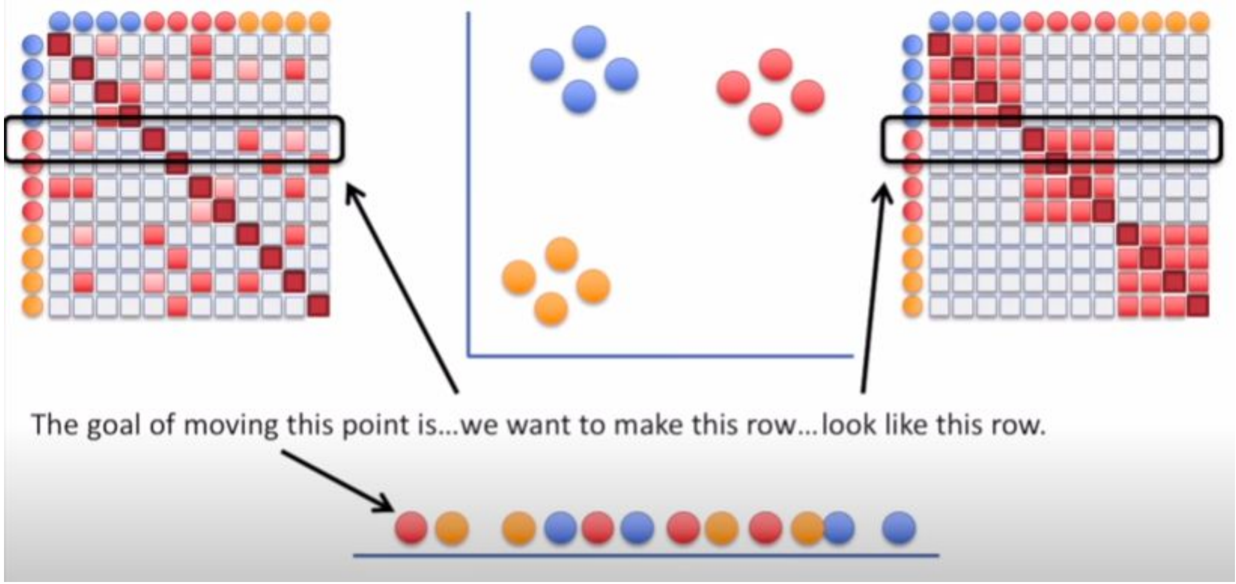
Findings by the authors:

- Low level features are learnt in the lower layers
- Many units learn the same concept
- ResNet has the highest number of Concept Detectors (*Bonus!*)
- Interpretability is independent of discriminative power and dependant on axis: Data was rotated and fed forward, the interpretability declined.



t-SNE

- A technique for visualizing high dimensional data
- It works by projecting the data to lower dimensions and clustering them such that they are similarly organized as in their high dimensional form.



So what now..

- Interpretability is possible
- (Holistic) Explainability is nearly impossible
- Keeping networks shallow can increase chances of holistic explainability

References and Further Reading

1. A big thank you to Christoph Molnar [https://christophm.github.io/]
2. Molnar's book on Interpretability in ML: https://christophm.github.io/book/
3. Network Dissection: http://netdissect.csail.mit.edu/
4. Interpretability in DL: https://distill.pub/2018/building-blocks/
5. Chakraborty et al. (2017): https://ieeexplore.ieee.org/document/8397411
6. Josh Starmer's Explanation of t-SNE: [https://www.youtube.com/watch?v=NEaUSP4YerM]