In Review of Network Dissection: Quantifying Interpretability of Deep Visual Representations

Emile Dos Santos Ferreira

6th February 2025

Department of Computer Science and Technology, University of Cambridge, England, United Kingdom

edsf2@cam.ac.uk

Context

- Networks can learn efficient encodings, using hidden variables to distinguish state
- Human-interpretable concepts are organically represented by the hidden units of DNNs [2]
- Ongoing research to encourage disentangled repr. [4]

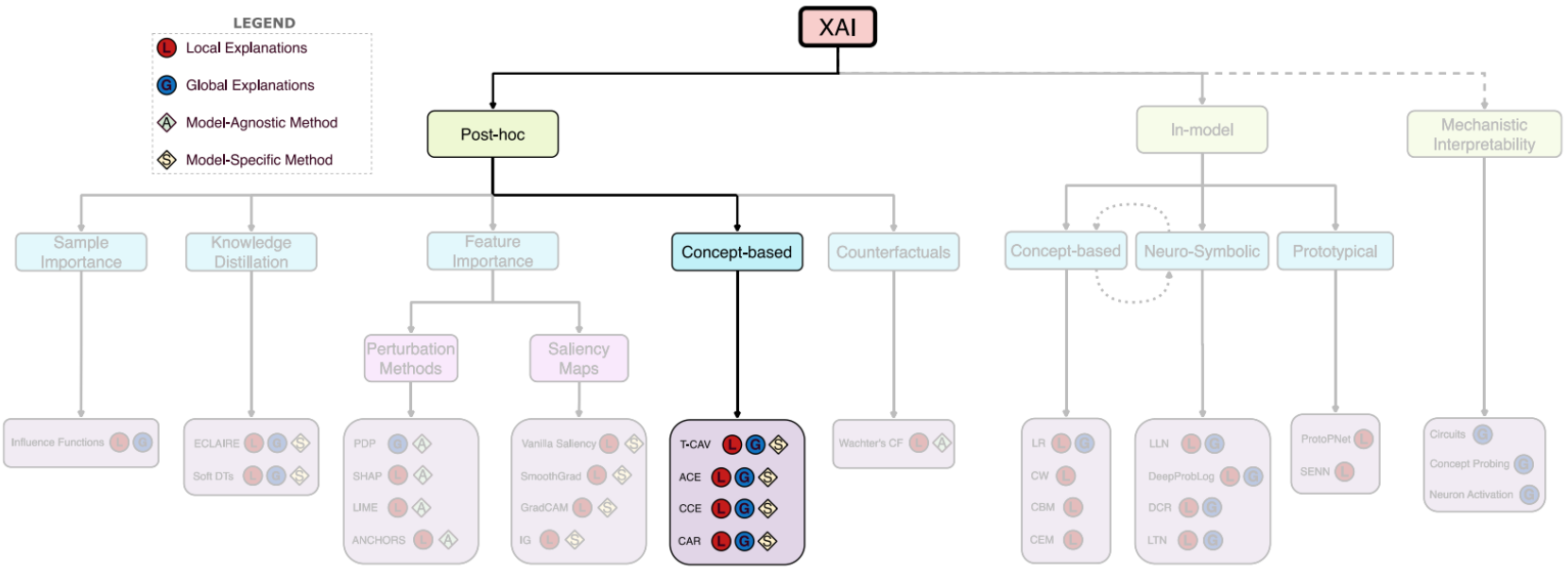
Questions

1. How can a disentangled representation be quantified?
2. Do interpretable units represent special alignment of feature space?
3. What training conditions lead to disentangled representations?

Approach

- Network Dissection: a framework for quantifying the interpretability of latent representations in CNNs
- Broden: a dataset of visual semantic concepts
- Evaluates alignment between hidden units and a set of semantic concepts to quantify their interpretability
- Investigate AlexNet, VGG, GoogLeNet and ResNet
- Experiment with axis rotation, different datasets and training techniques

Taxonomy


LEGEND
Local Explanations
Global Explanations
Model-Agnostic Method
Model-Specific Method
XAI
Post-hoc
In-model
Mechanistic Interpretability
Sample Importance
Knowledge Distillation
Feature Importance
Concept-based
Counterfactuals
Concept-based
Neuro-Symbolic
Prototypical
Perturbation Methods
Saliency Maps
Influence Functions
ECLAIRE
Soft DTs
PDP
SHAP
LIME
ANCHORS
Vanilla Saliency
SmoothGrad
GradCAM
IG
T-CAV
ACE
CCE
CAR
Wachter's CF
LR
CW
CBM
CEM
LLN
DeepProbLog
DCR
LTN
ProtoPNet
SENN
Circuits
Concept Probing
Neuron Activation

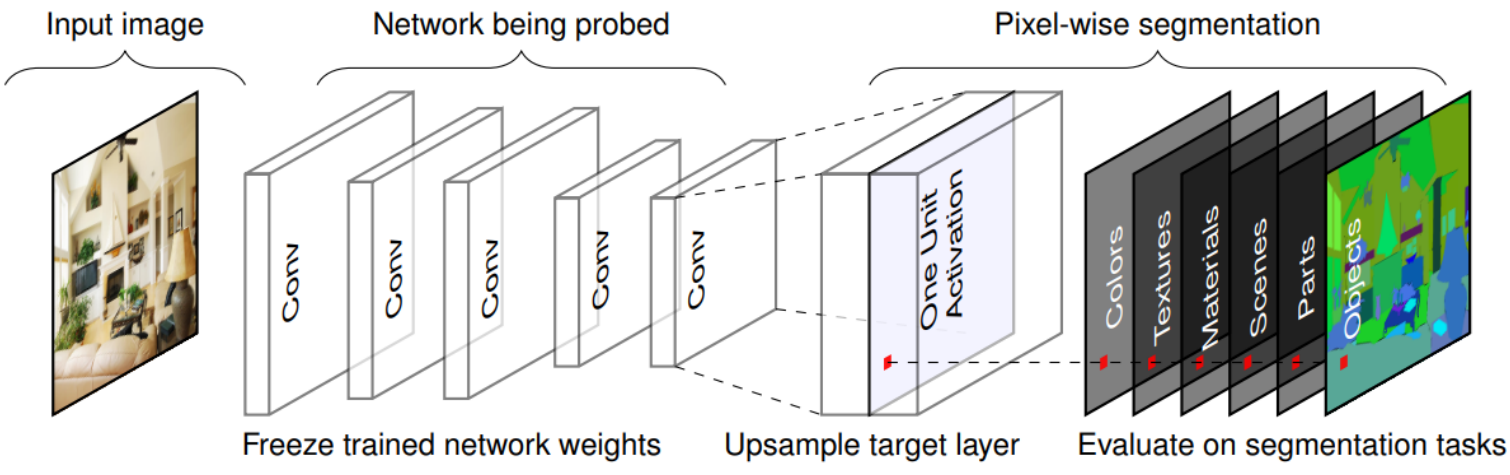
Related work

There existed tools for interpreting CNNs in 2017, but none quantitative.

- Deconvolution: sampling image patches to maximise activations [9]
- Back-propagation: creating saliency maps [8]
- Visualising activations and receptive fields for concepts [6]

Other related work includes training linear probes for intermediate layers [1] and generating prototypical images from feature inversion mapping [7].

Network Dissection



Annotated dataset

- **Bro**adly and **den**sely (Broden) annotated dataset
- Totalling 63,305 images and 1,197 visual concepts
- Each pixel can have multiple labels
- Labels have at least 10 image samples

Category	Classes	Sources	Avg sample
scene	468	ADE [43]	38
object	584	ADE [43], Pascal-Context [19]	491
part	234	ADE [43], Pascal-Part [6]	854
material	32	OpenSurfaces [4]	1,703
texture	47	DTD [7]	140
color	11	Generated	59,250

Intersection over union (IoU) score

- Given $A_k(\mathbf{x})$, determine T_k s.t. $P(a_k > T_k) = 0.005$
- Scale $A_k(\mathbf{x})$ using bilinear interpolation to $S_k(\mathbf{x})$
- Threshold to binary classification $M_k(\mathbf{x}) \equiv S_k(\mathbf{x}) \geq T_k$
- Equation 1 defines accuracy of unit k in detecting c

$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|} \tag{1}$$

Ts and Cs

- Computed only on the images that contain a concept of the category of c
- It does not measure discriminative power
- Unit is a detector of c if $IoU_{k,c} > 0.04$
- Choose label with highest score in case of multiple
- Measures alignment between a single unit and concept
- Quantify a layer by summing over its unique detectors

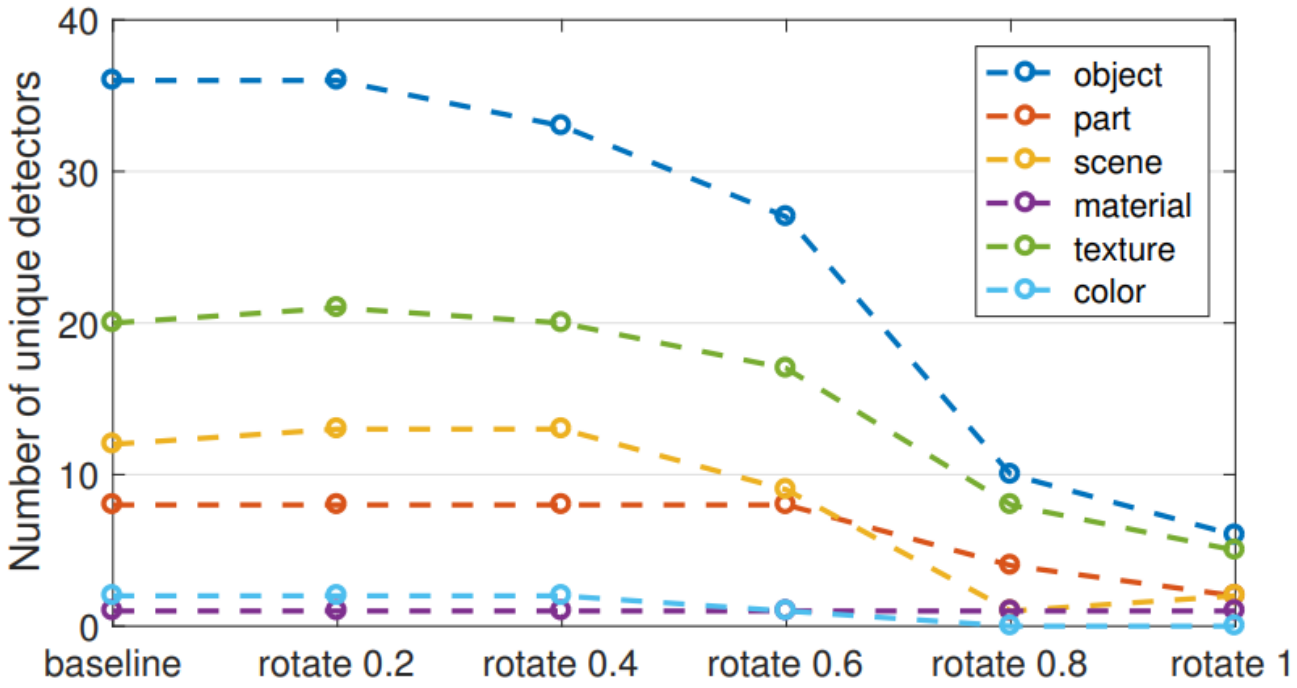
Experiment: human evaluation

- Some number of Amazon Mechanical Turk workers
- 15 images with highlighted patches of high activations
- AlexNet trained on Places205
- Asked if a given phrase describes most of the patches

	conv1	conv2	conv3	conv4	conv5
Interpretable units	57/96	126/256	247/384	258/384	194/256
Human consistency	82%	76%	83%	82%	91%
Network Dissection	37%	56%	54%	59%	71%

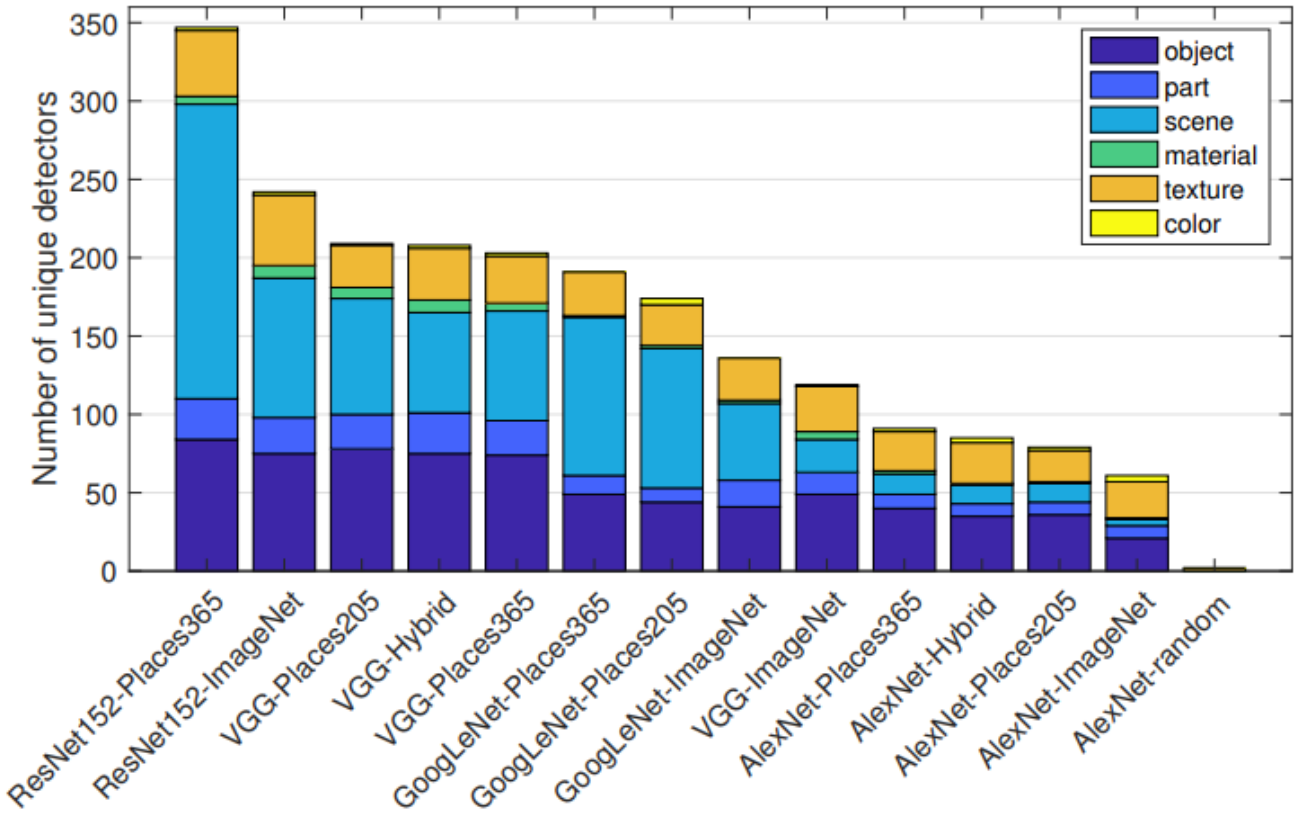
Experiment: axis rotation

- Same model and dataset as before
- Apply rotation matrix to the model output
- Discriminative power remains constant
- Reduced # unique detectors by 80%
- Implies special alignment of feature space



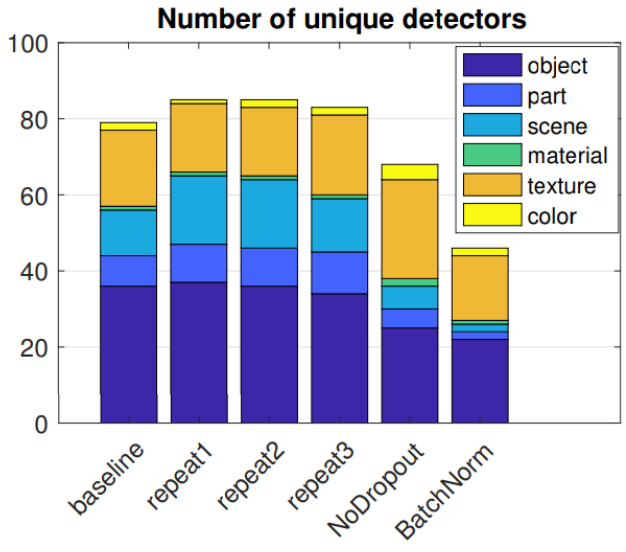
Experiment: model and dataset

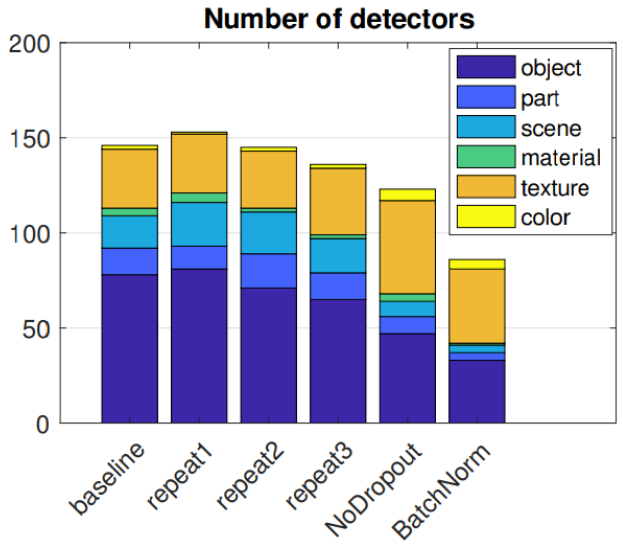
- # layers $\propto$ # unique detectors $\propto$ IoU
- Difference in # scenes



Experiment: regularisation

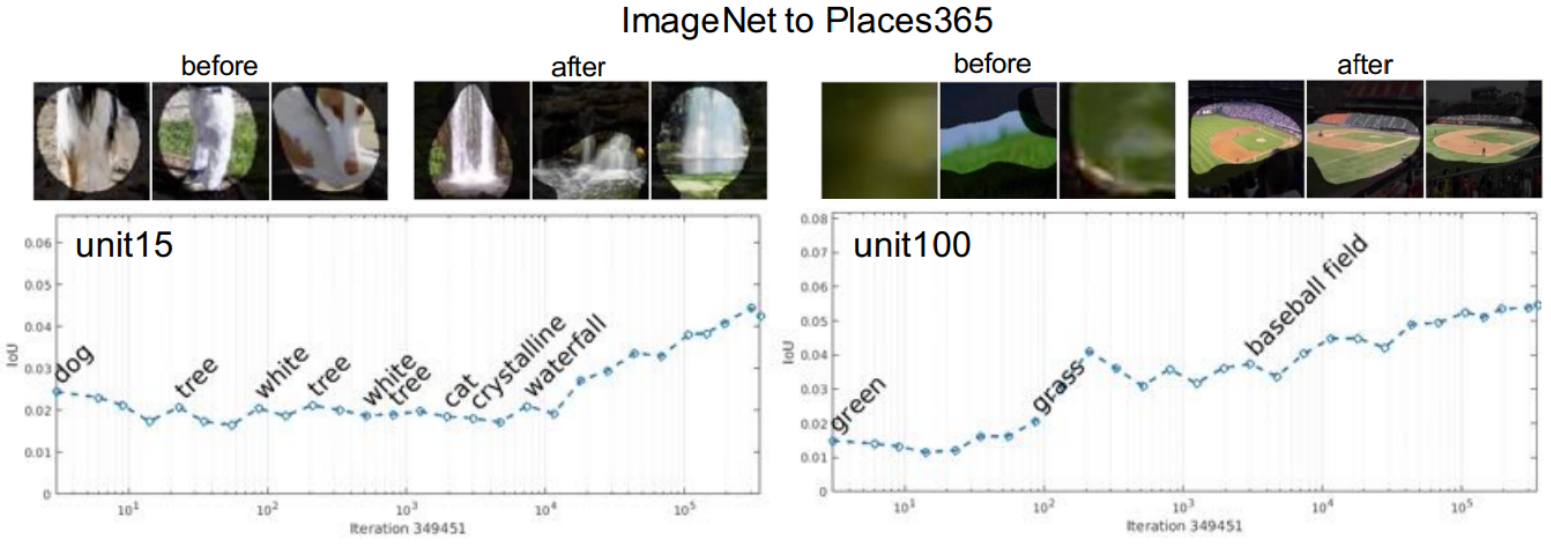
- NoDropout: remove random dropout from FC layers
- BatchNorm: add batch norm. to all conv. layers





Experiment: fine-tuning (extended paper)

- Change a unit's concept to a visually similar one



Conclusion

1. How can a disentangled representation be quantified? Using the IoU score of the Network Dissection framework
2. Do interpretable units represent special alignment of feature space? Yes, interpretability is not axis-independent
3. What training conditions lead to disentangled representations? Findings:
   - Interpretability $\propto$ model depth
   - Discriminative power $\not\propto$ interpretability
   - BatchNorm reduces interpretability

Review

Positive

- Supports any CNN and does not require back-propagation
- Provide code and data (fittingly) under MIT License
- Website[1] with results, visualisations and videos
- Released an optimised library a year later

Negative

- 94% of dataset made up by $< 1\%$ of classes
- A threshold accuracy of 4% is very low
- Use reference numbers as nouns in related work
- No future work section
- Human evaluation: only 15 images and ? participants

[1] `netdissect.csail.mit.edu`

Impact

- First method to quantify the interpretability of CNNs
- 1800+ citations and 440+ stars on GitHub
- Broden dataset widely used: Net2Vec [5]

Future work

- Followed up with a paper on *Understanding the Role of Individual Units in a Deep Network* [3]
- Increase the dataset of 1,197 visual concepts, as many units unidentified
- Enable batch normalisation to preserve interpretability

References i


G. Alain and Y. Bengio.
Understanding intermediate layers using linear classifier probes, 2018.


D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba.
Network dissection: Quantifying interpretability of deep visual representations.
In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

References ii

D. Bau, J.-Y. Zhu, H. Strobelt, A. Lapedriza, B. Zhou, and A. Torralba.
Understanding the role of individual units in a deep neural network.
Proceedings of the National Academy of Sciences, 2020.

Y. Bengio, A. Courville, and P. Vincent.
Representation learning: A review and new perspectives.
IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

References iii

R. Fong and A. Vedaldi.
Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

R. Girshick, J. Donahue, T. Darrell, and J. Malik.
Rich feature hierarchies for accurate object detection and semantic segmentation.
In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

References iv

- A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox, and J. Clune.
  Synthesizing the preferred inputs for neurons in neural networks via deep generator networks.
  Advances in neural information processing systems, 29, 2016.
- K. Simonyan.
  Very deep convolutional networks for large-scale image recognition.
  arXiv preprint arXiv:1409.1556, 2014.


M. D. Zeiler and R. Fergus.
Visualizing and understanding convolutional networks.
In D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing.

Experiment: training duration

- # unique detectors increases with training
- Good proxy for validation accuracy (extended paper)