Network Dissection: Quantifying Interpretability of Deep Visual Representations

David Bau*, Bolei Zhou*, Aditya Khosla, Aude Oliva, Antonio Torralba

* Indicates equal contribution

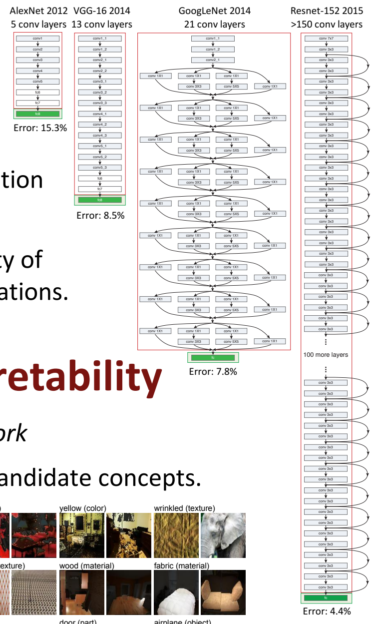
Peering In

What is learned inside?
How do internals compare?

Our contributions:

1. A method to go from visualization to quantified interpretations.
2. Comparisons of interpretability of a range of different representations.

Output measured by performance.
Internal units match unknown concepts.



Quantifying Interpretability

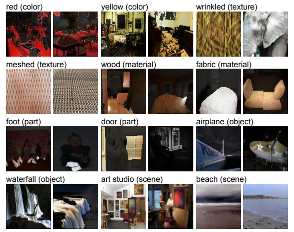
The Network Dissection Framework

1. Define a broad dictionary of candidate concepts.

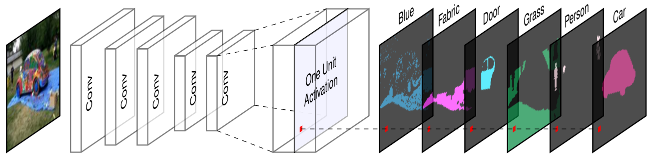
Broden Dataset

ADE20K	Zhou et al, CVPR '17
Pascal Context	Mottaghi et al, CVPR '14
Pascal Part	Chen et al, CVPR '14
Open Surfaces	Bell et al, SIGGRAPH '14
Desc Textures	Cimpoi et al, CVPR '14
Colors	generated

Total = 63,305 images
1,197 concepts



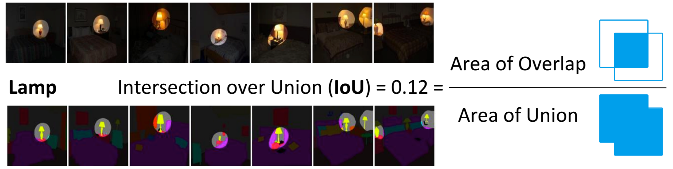
2. Test each internal unit on segmentation of every concept.



3. Measure segmentation quality and match units to concepts.

Unit 2 Top activated image areas

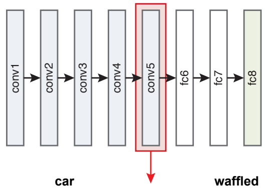
Lamp Intersection over Union (IoU) = 0.12 = Area of Overlap / Area of Union



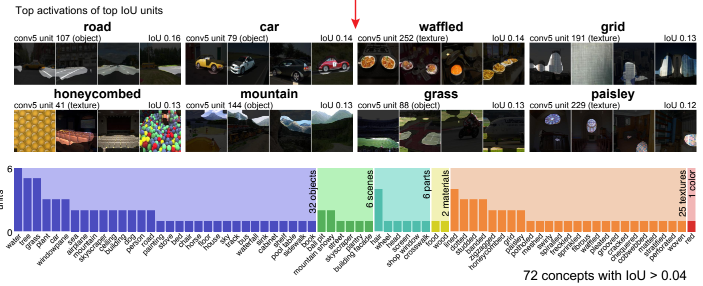
IoU of the best-matched concepts quantify interpretability

Are Individual Units Meaningful?

1. Dissect 256 units of Alexnet conv5 trained on places

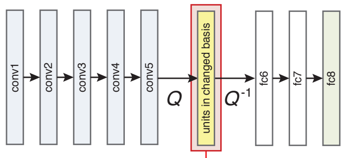


Top ranked concept and score are assigned to each unit.

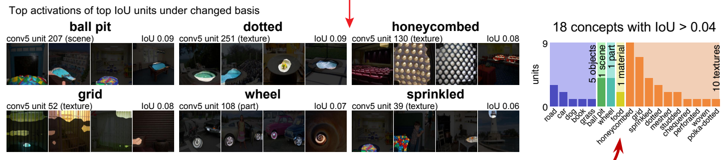


2. Dissect 256 other projections of the same Alexnet conv5 units

Representation after random basis-change has identical discriminative power

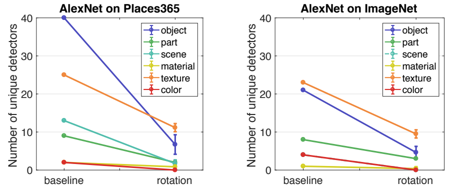


Top ranked concept and score are assigned to each unit.



Units under a changed basis are less interpretable

3. Verify on other projections

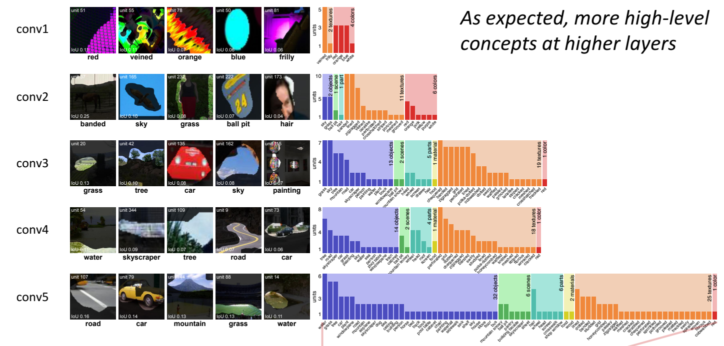


Learned units match Interpretations better than a typical basis.

Individual units in a learned basis match meaningful concepts

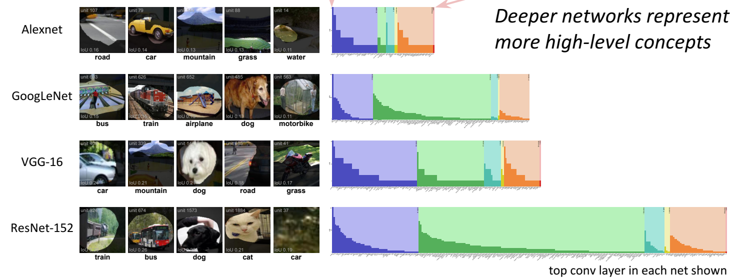
Comparing Interpretable Units

1. Across Layers (Alexnet trained on places)



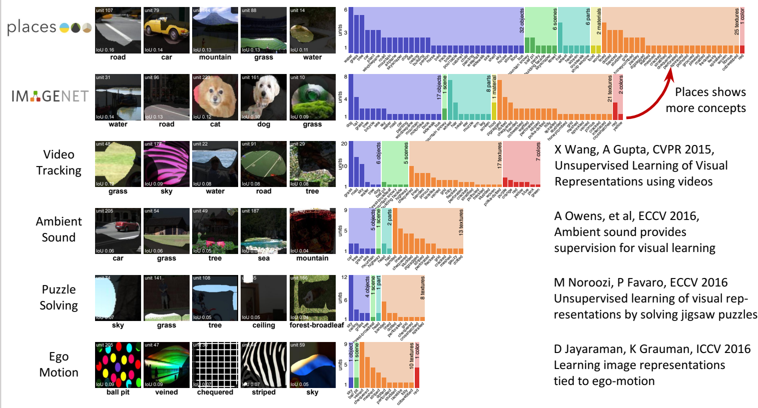
As expected, more high-level concepts at higher layers

2. Across Architectures (trained on places)



Deeper networks represent more high-level concepts

3. Across Supervisions and Training Sets (Alexnet conv5)



X Wang, A Gupta, CVPR 2015, Unsupervised Learning of Visual Representations using videos

A Owens, et al, ECCV 2016, Ambient sound provides supervision for visual learning

M Noroozi, P Favaro, ECCV 2016 Unsupervised learning of visual representations by solving jigsaw puzzles

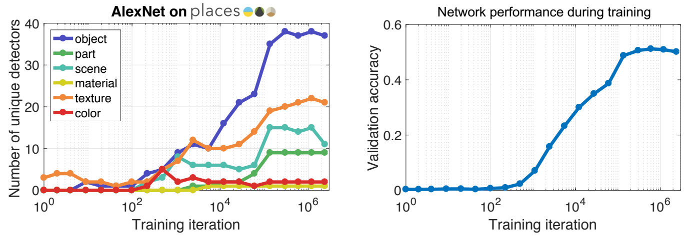
D Jayaraman, K Grauman, ICCV 2016 Learning image representations tied to ego-motion

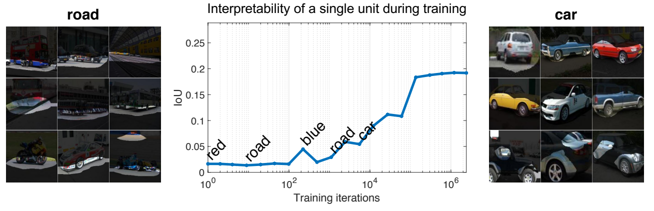
Representations can be compared by interpretability

Emergence of Interpretability

1. When Training from Scratch

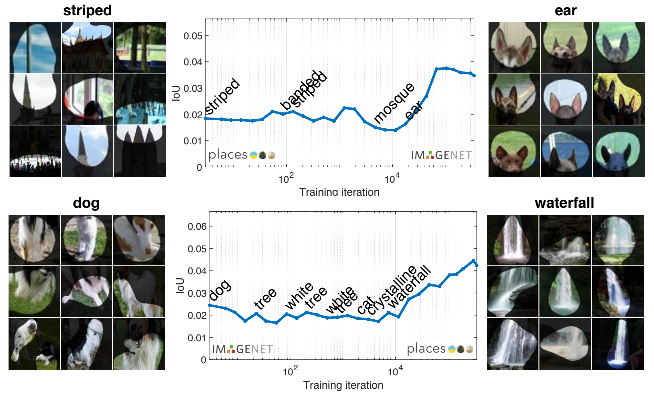
Early training finds concepts; late training improves them.





2. When Fine-Tuning

Representations switch units to new concepts during fine-tuning.



Papers, data, and code at http://netdissect.csail.mit.edu

Acknowledgements: This work was partly supported by the National Science Foundation under Grant No. 1524817 to A.T.; the Vannevar Bush Faculty Fellowship program sponsored by the Basic Research Office of the Assistant Secretary of Defense for Research and Engineering and funded by the Office of Naval Research through grant N00014-16-1-3116 to A.O.; the MIT Big Data Initiative at CSAIL, the Toyota Research Institute MIT CSAIL Joint Research Center, Google and Amazon Awards, and a hardware donation from NVIDIA Corporation. B.Z. is supported by a Facebook Fellowship.