Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks

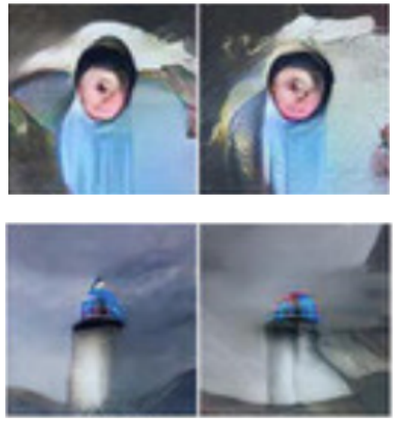
Ruth Fong and Andrea Vedaldi

UNIVERSITY OF OXFORD

Filter Visualizations

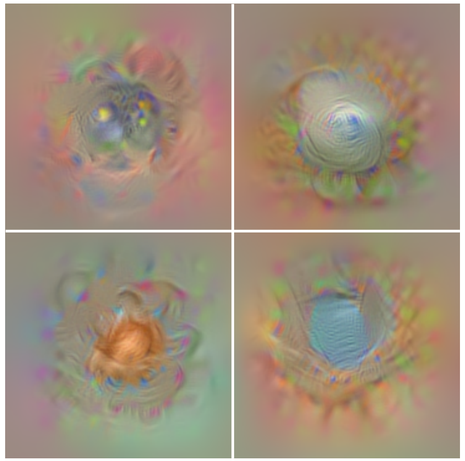
Zeiler & Fergus, ECCV 2014

Nyugen et al., NIPS 2016



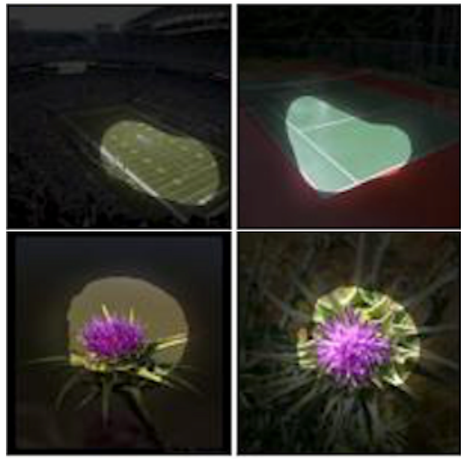
Zhou et al., ICLR 2015

Mahendran and Vedaldi, IJCV 2016

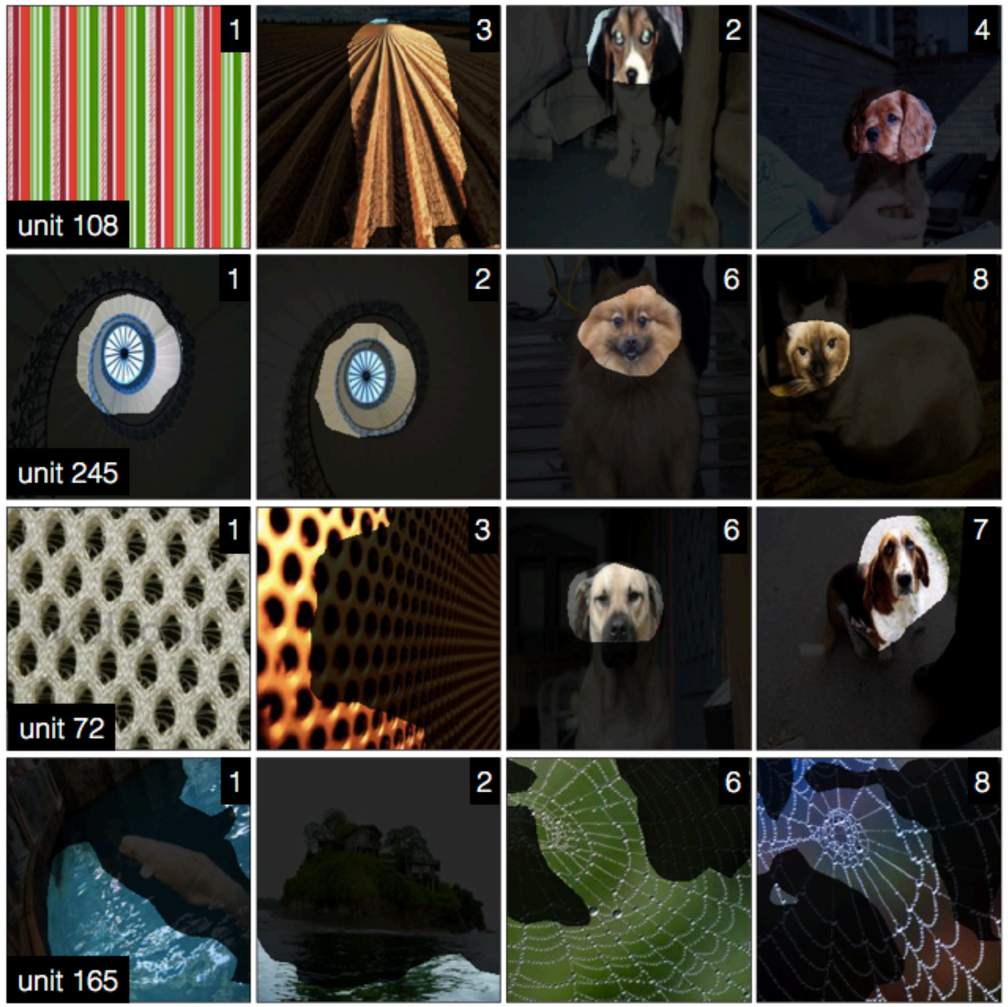


Olah et al., Distill 2017

Bau et al., CVPR 2017

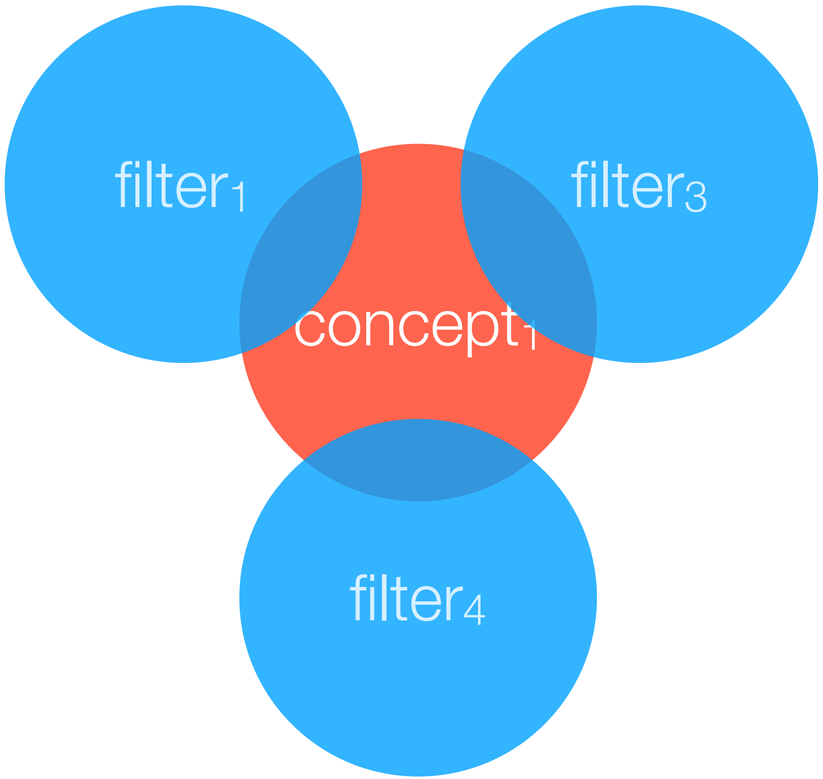


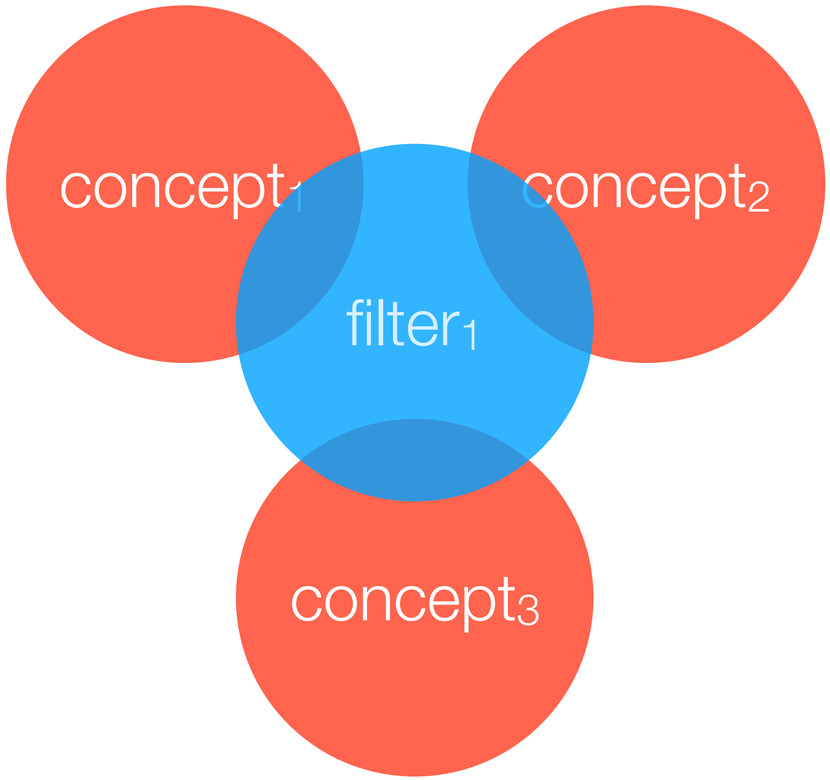
Filter-Concept Overlap



Modified Network Dissection visualization of AlexNet conv5 filters [Bau, et al., 2017]

Filter-Concept Overlap





Method

Probe a **network** with a **dataset** and learn to perform a **task** using **activations** at a given layer.

Bau et al, CVPR 2017. “Network Dissection.”
Agrawal et al., ECCV 2014. “Analyzing the performance of neural networks.”
Alain & Bengio, ICLR Workshop 2017. “Understanding intermediate layers using linear probes.”

BRODEN Dataset

Image-level Annotations **Pixel-level Annotations**

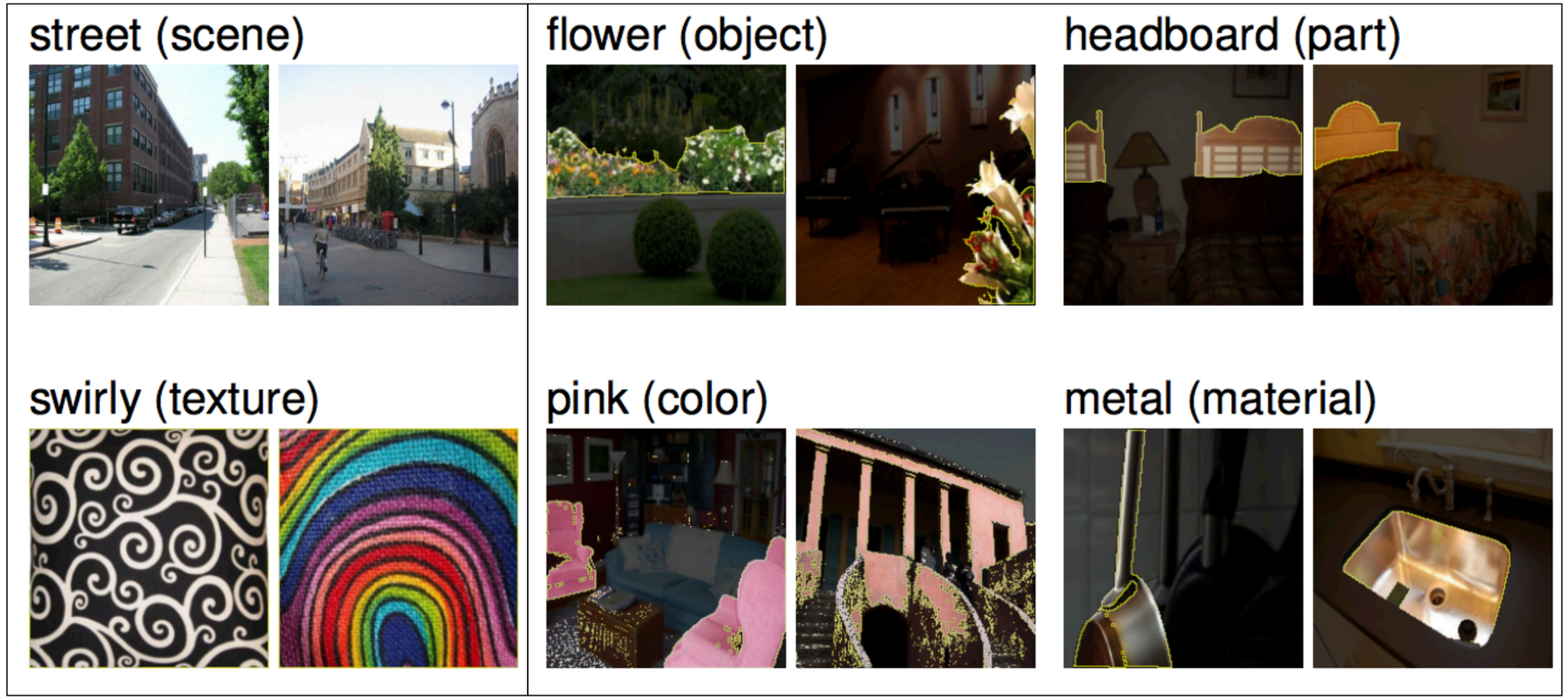
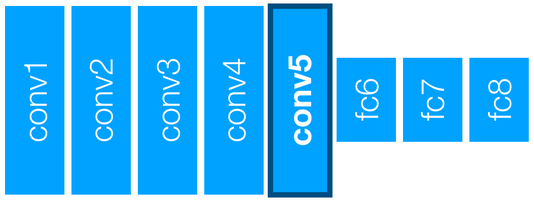


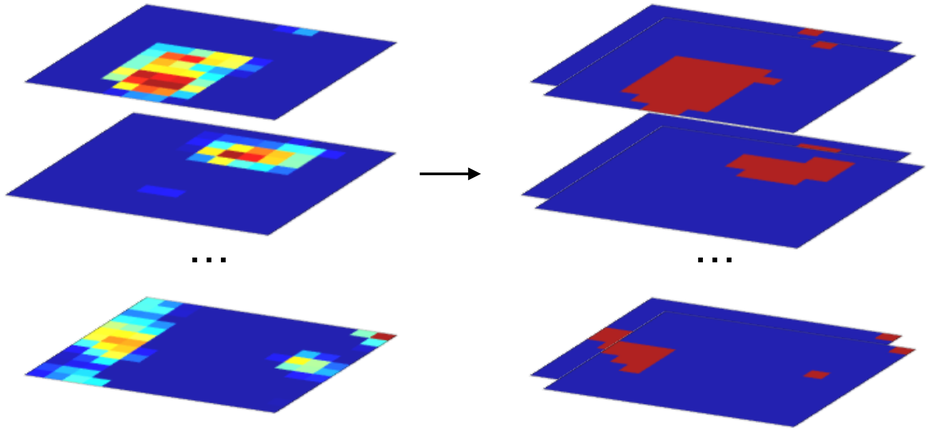
Figure 1 from Bau, et al., 2017

Segmentation

Extract Activations



Threshold Activations



99.5% quantile
as in Bau et al., 2017

Linearly Combine Activations

$\times\ w_1$

$\times\ w_2$

...

$\times\ w_K$

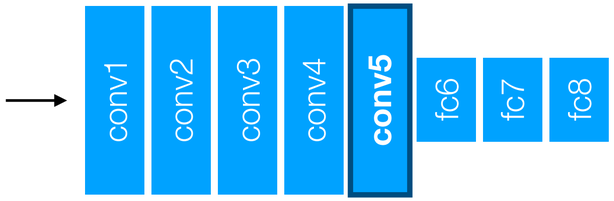
$\sum$

Segmentation Mask

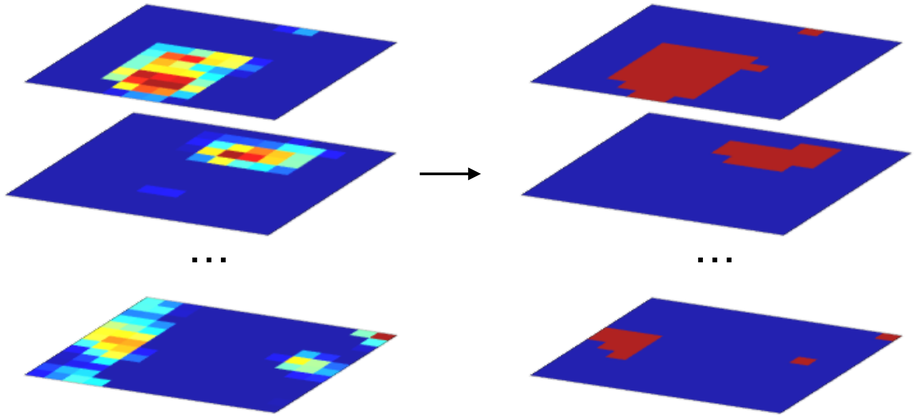
IoU = .77

Segmentation

Extract Activations



Threshold Activations



99.5% quantile as in Bau et al., 2017

Linearly Combine Activations

$$\times w_1 \quad \times w_2 \quad \dots \quad \times w_K \rightarrow \sum$$

Concept Vector

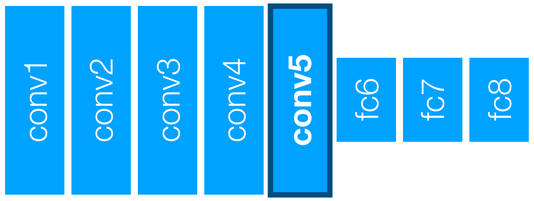
$\mathbf{w}_{dog}$

Segmentation Mask

IoU = .77

Segmentation

Extract Activations



$$\mathbf{w}_{dog,F} = \begin{bmatrix} w_1 \\ \dots \\ w_F \end{bmatrix}$$

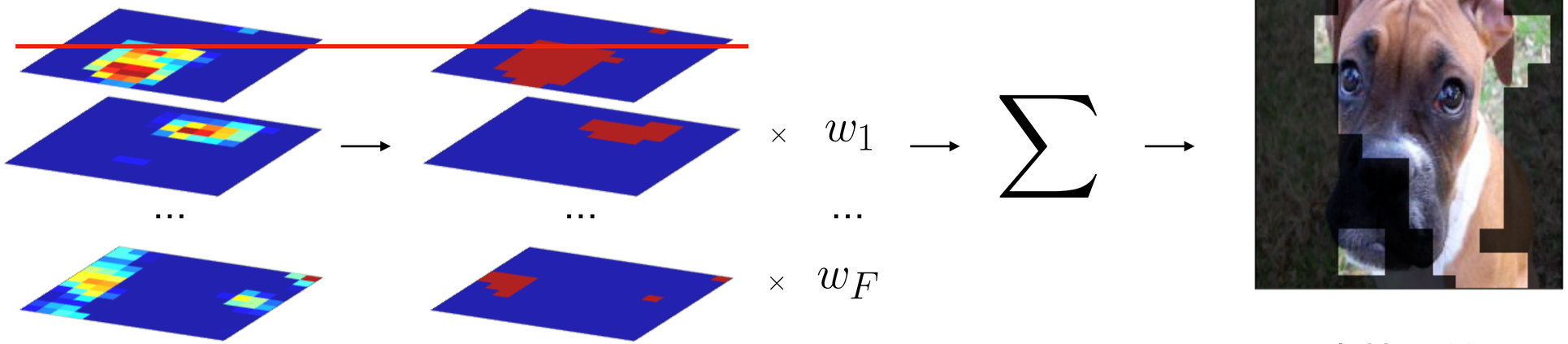
Subset:
Only use top *F* filters, chosen by magnitude (*F* = 4)

Subset selection follows Agrawal et al., 2014

Threshold Activations

Linearly Combine Activations

Segmentation Mask

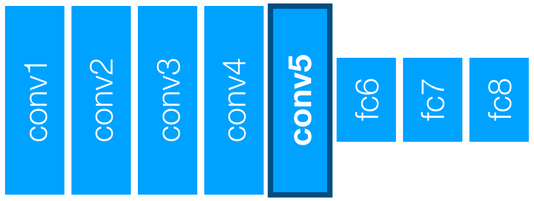


IoU = .63

Segmentation

Single Filter

Extract Activations

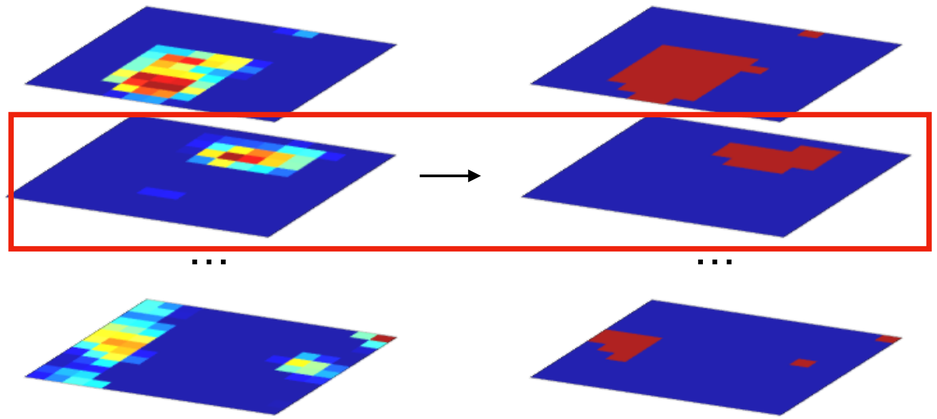


$$\mathrm{IoU}_{\mathrm{set}}(c; M, s) = \frac{\sum_{\mathbf{x} \in X_{s,c}} |M(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum_{\mathbf{x} \in X_{s,c}} |M(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

$$\mathrm{IoU}_{\mathrm{ind}}(\mathbf{x}, c; M) = \frac{|M(\mathbf{x}) \cap L_c(\mathbf{x})|}{|M(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

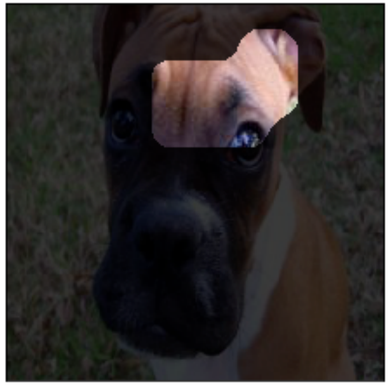
Threshold Activations

Filter 169



Choose Best Filter

Segmentation Mask

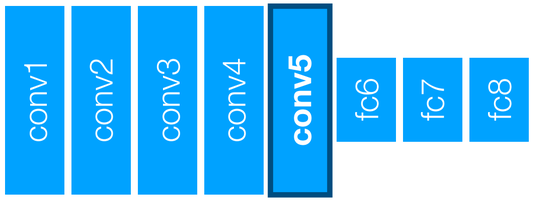


IoU = .18

Near equivalent to Bau et al., 2017

Classification

Extract Activations

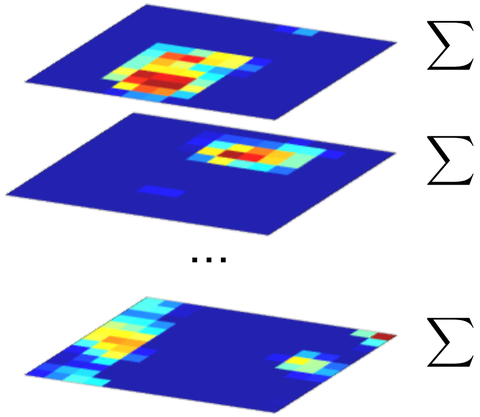


$$\mathbf{w}_{dog} = \begin{bmatrix} w_1 \\ w_2 \\ \dots \\ w_K \end{bmatrix}$$

Global Average Pooling

Linearly Combine Activations (+ bias)

Classification Prediction



$$\sum \quad \bar{a}_1 \times w_1$$

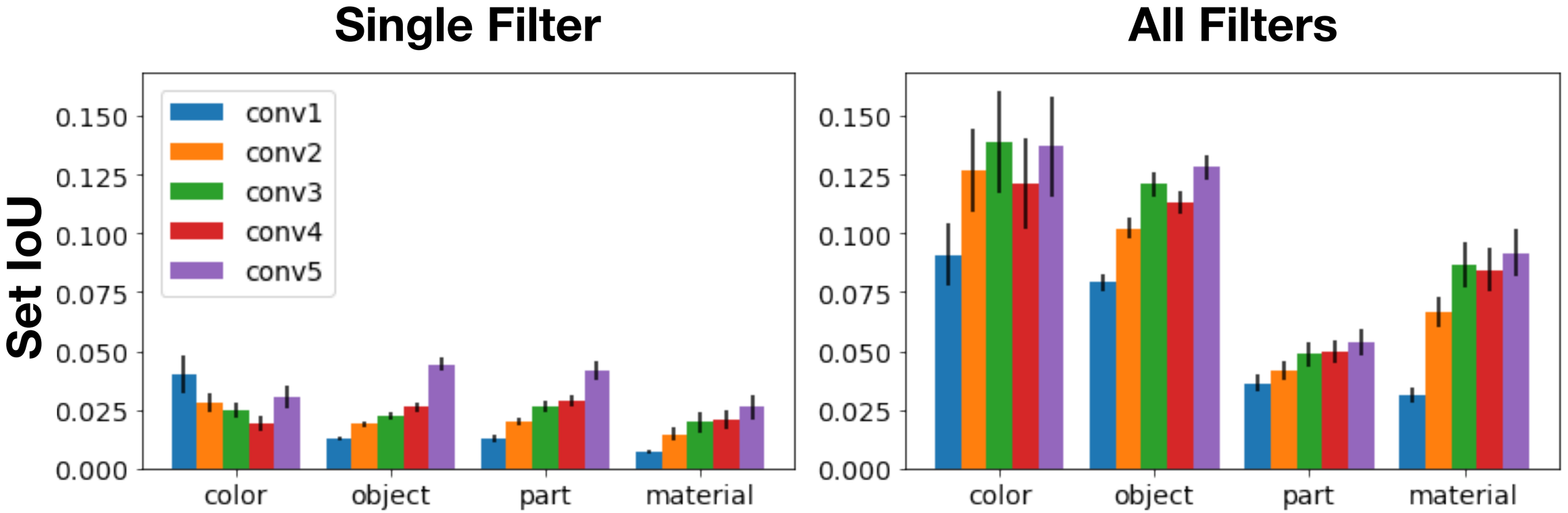
$$\sum \quad \bar{a}_2 \times w_2 \rightarrow \sum_{+b} \rightarrow \{\text{dog}, \text{no dog}\}$$

$$\dots \qquad \dots$$

$$\sum \quad \bar{a}_K \times w_K$$

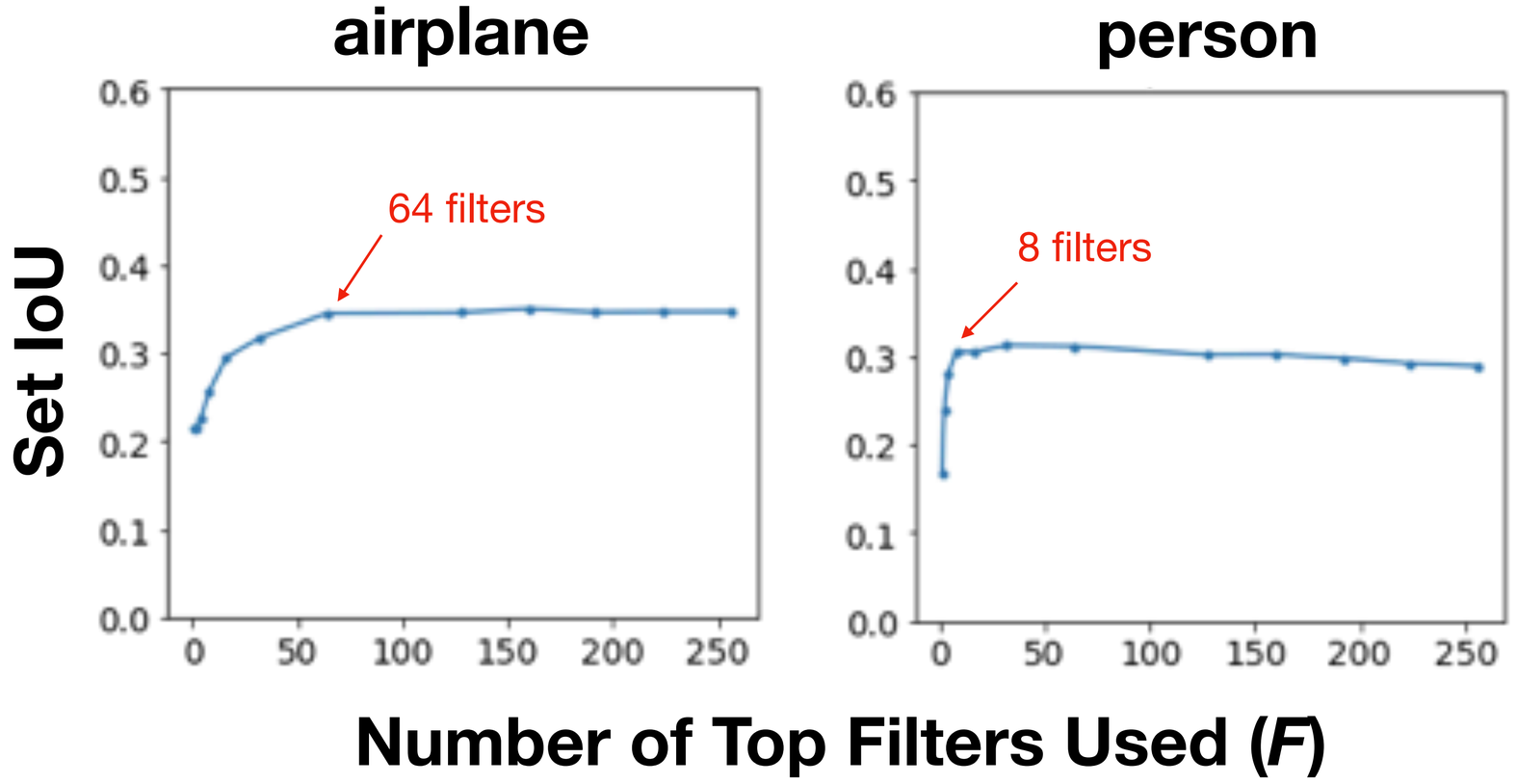
A Few Results

Single vs. All Filters



Concepts are encoded better when using multiple filters.

Filters Per Concept



Different concepts require different number of filters for encoding.

Filters: Supervised vs. Self-Supervised

Performance Improvement (Single Filter → All Filters):

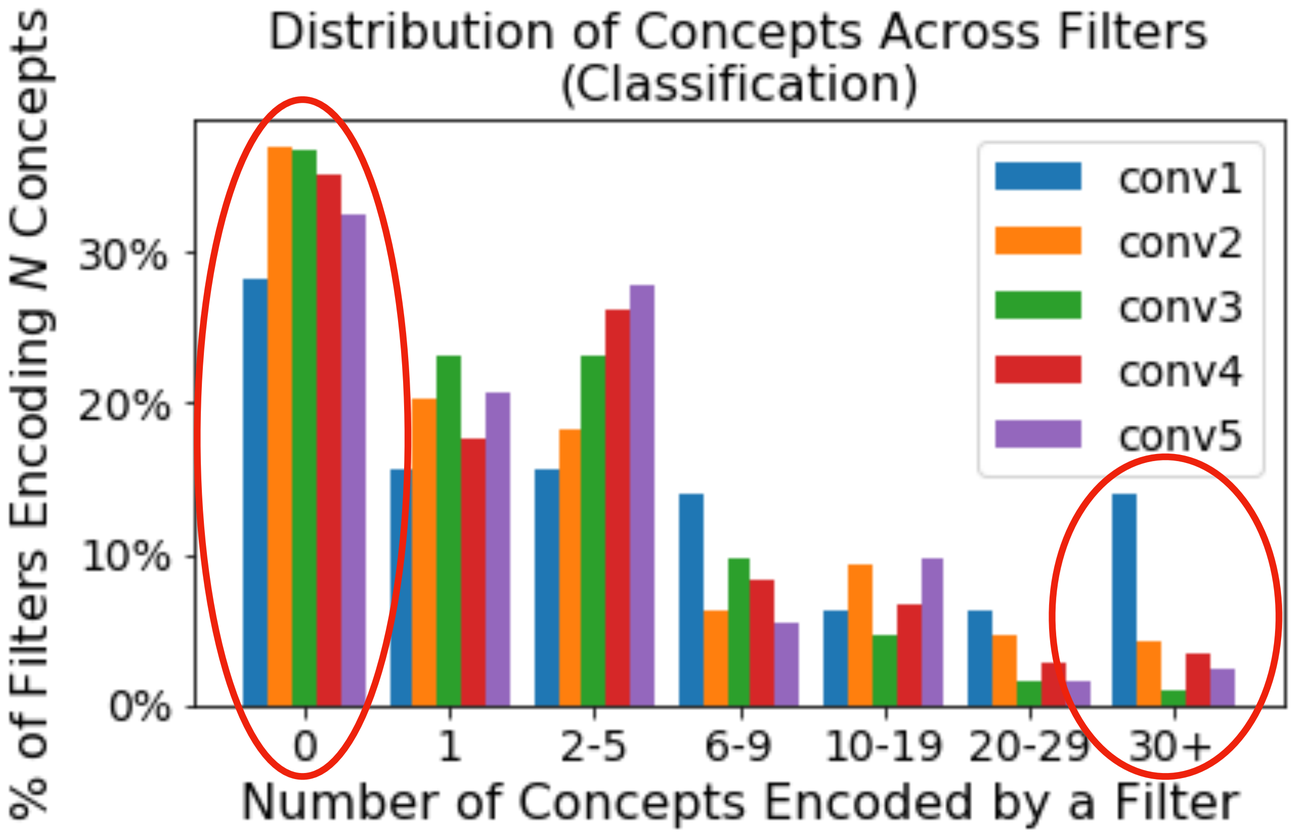
- Self-supervised networks: 5-6x
- Fully-supervised networks: 2-4x

Self-supervised networks encode BRODEN concepts more distributively.

Concepts per Filters

Found a wide range in filter capacity to encode concepts:

- Many filters aren't selective for any concepts
- A few filters are selective for many concepts



Concepts per Filters

AlexNet conv5 unit 66 is highly selective for various farm animals

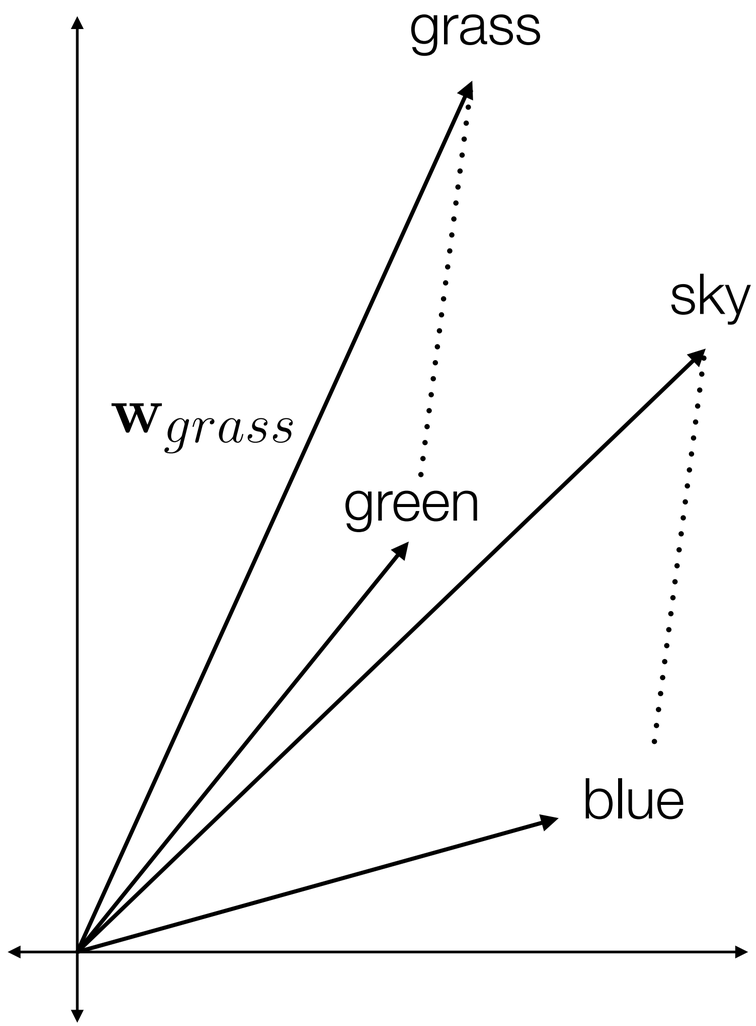
Sheep (IoU_{set} = .21)

Horse (IoU_{set} = .21)

Cow (IoU_{set} = .20)

Visualizing Non-Maximal Examples

Comparing Concept Embeddings

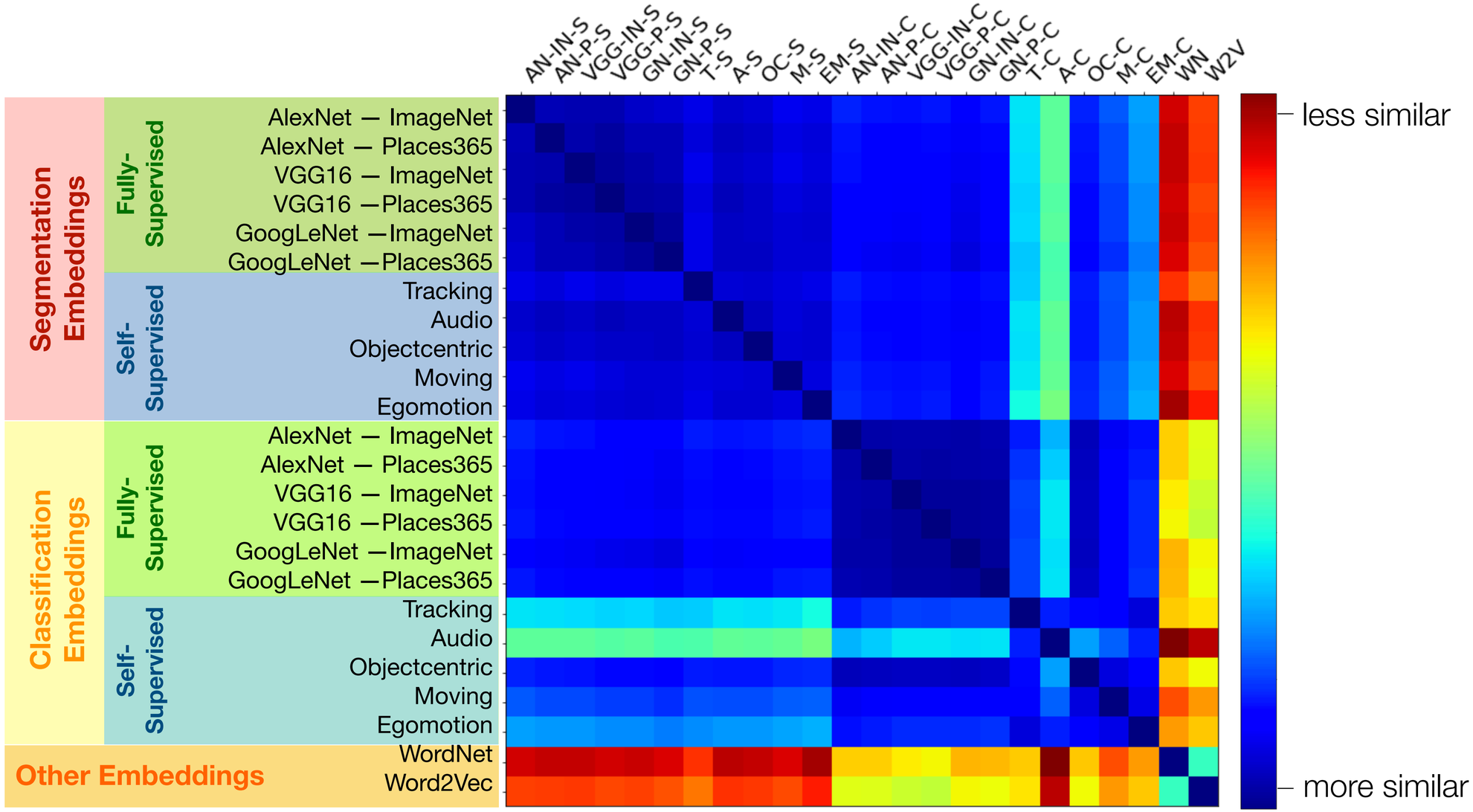


grass + blue — green = sky

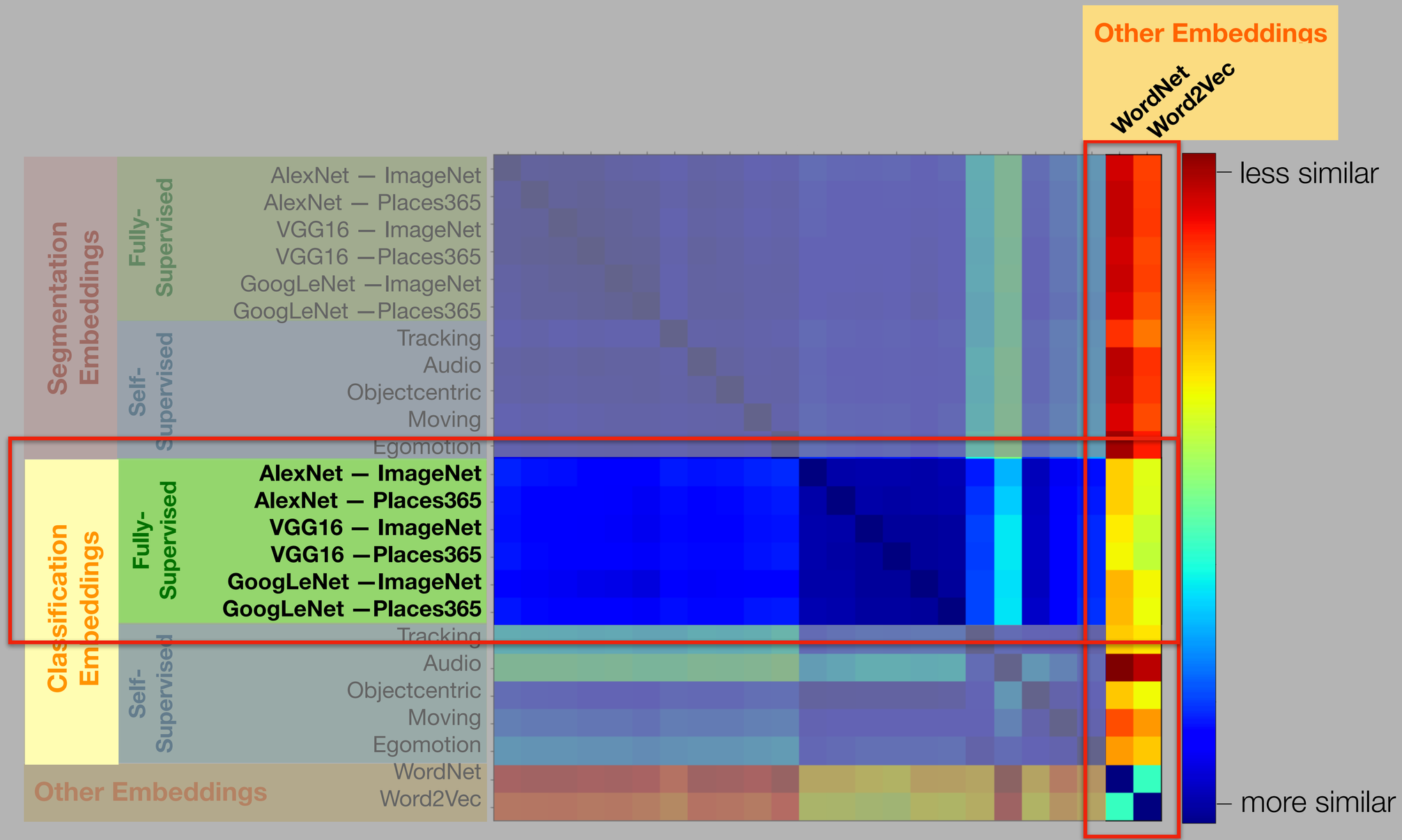
tree — wood = plant

person — torso = foot

Comparing Concept Embedding **Spaces**



Comparing Concept Embedding **Spaces**



Chat more at poster E9!

Code: https://github.com/ruthcfong/net2vec