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

Ruth Fong and Andrea Vedaldi
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.

BRODEN Dataset

Image-level Annotations
- street (scene)
- swirly (texture)

Pixel-level Annotations
- flower (object)
- pink (color)
- headboard (part)
- metal (material)

Figure 1 from Bau, et al., 2017
**Segmentation**

99.5% quantile as in Bau et al., 2017

Threshold Activations

Linearly Combine Activations

$\sum \times w_1 \times w_2 \times w_K$

Segmentation Mask

IoU = .77
Segmentation

Extract Activations

Conv1 → Conv2 → Conv3 → Conv4

Threshold Activations

Conv5

fc6 → fc7 → fc8

Linearly Combine Activations

\[ \sum \]

Linear Combination

IoU = .77

Concept Vector

\[ w_{dog} \]

99.5% quantile as in Bau et al., 2017
**Segmentation**

**Extract Activations**

\[
\begin{bmatrix}
  w_1 \\
  \vdots \\
  w_F
\end{bmatrix}
\]

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

Subset selection follows Agrawal et al., 2014

**Threshold Activations**

**Linearly Combine Activations**

\[
\sum \times w_1 
\]

**Segmentation Mask**

IoU = .63
Segmentation

Extract Activations

Filter 169

Threshold Activations

Choose Best Filter

Segmentation Mask

Near equivalent to Bau et al., 2017

Single Filter

\[
\text{IoU}_{\text{set}}(c; M, s) = \frac{\sum_{x \in X_{s,c}} |M(x) \cap L_c(x)|}{\sum_{x \in X_{s,c}} |M(x) \cup L_c(x)|}
\]

\[
\text{IoU}_{\text{ind}}(x, c; M) = \frac{|M(x) \cap L_c(x)|}{|M(x) \cup L_c(x)|}
\]

IoU = .18
Classification

Extract Activations

\[ \text{conv1, conv2, conv3, conv4, conv5, fc6, fc7, fc8} \]

Global Average Pooling

\[ \sum a_1 \times w_1 \]
\[ \sum a_2 \times w_2 \]
\[ ... \]
\[ \sum a_K \times w_K \]

Linearly Combine Activations (+ bias)

\[ \mathbf{w}_{\text{dog}} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_K \end{bmatrix} \]

Classification Prediction

\[ \sum + b \]

\{dog, no dog\}
A Few Results
Single vs. All Filters

Concepts are encoded better when using multiple filters.
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.
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
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

All Filters

IoU

set

= .35

our

Best Filter

IoU

set

= .14

Bau et al., 2017

Spanning Deciles

10% 20% 30% 40% 50% 60% 70% 80% 90%

.01 .02 .24 .28 .33 .37 .39 .43 .49

.02 .05 .10 .16 .19 .27 .31 .37 .46

.01 .02 .24 .28 .33 .37 .39 .43 .49
Comparing Concept Embeddings

\[ \mathbf{w}_{\text{grass}} \]

- grass + blue $-$ green = sky
- tree $-$ wood = plant
- person $-$ torso = foot
Comparing Concept Embedding Spaces

- less similar
- more similar
Comparing Concept Embedding Spaces

- Segmentation Embeddings
- Fully-Supervised
- Self-Supervised
- AlexNet — ImageNet
- AlexNet — Places365
- VGG16 — ImageNet
- VGG16 — Places365
- GoogLeNet — ImageNet
- GoogLeNet — Places365
- Tracking
- Audio
- Objectcentric
- Moving
- Egomotion

- Classification Embeddings
- Fully-Supervised
- Self-Supervised
- AlexNet — ImageNet
- AlexNet — Places365
- VGG16 — ImageNet
- VGG16 — Places365
- GoogLeNet — ImageNet
- GoogLeNet — Places365
- Tracking
- Audio
- Objectcentric
- Moving
- Egomotion

- Other Embeddings
- WordNet
- Word2Vec
- less similar
- more similar
Chat more at poster E9!

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