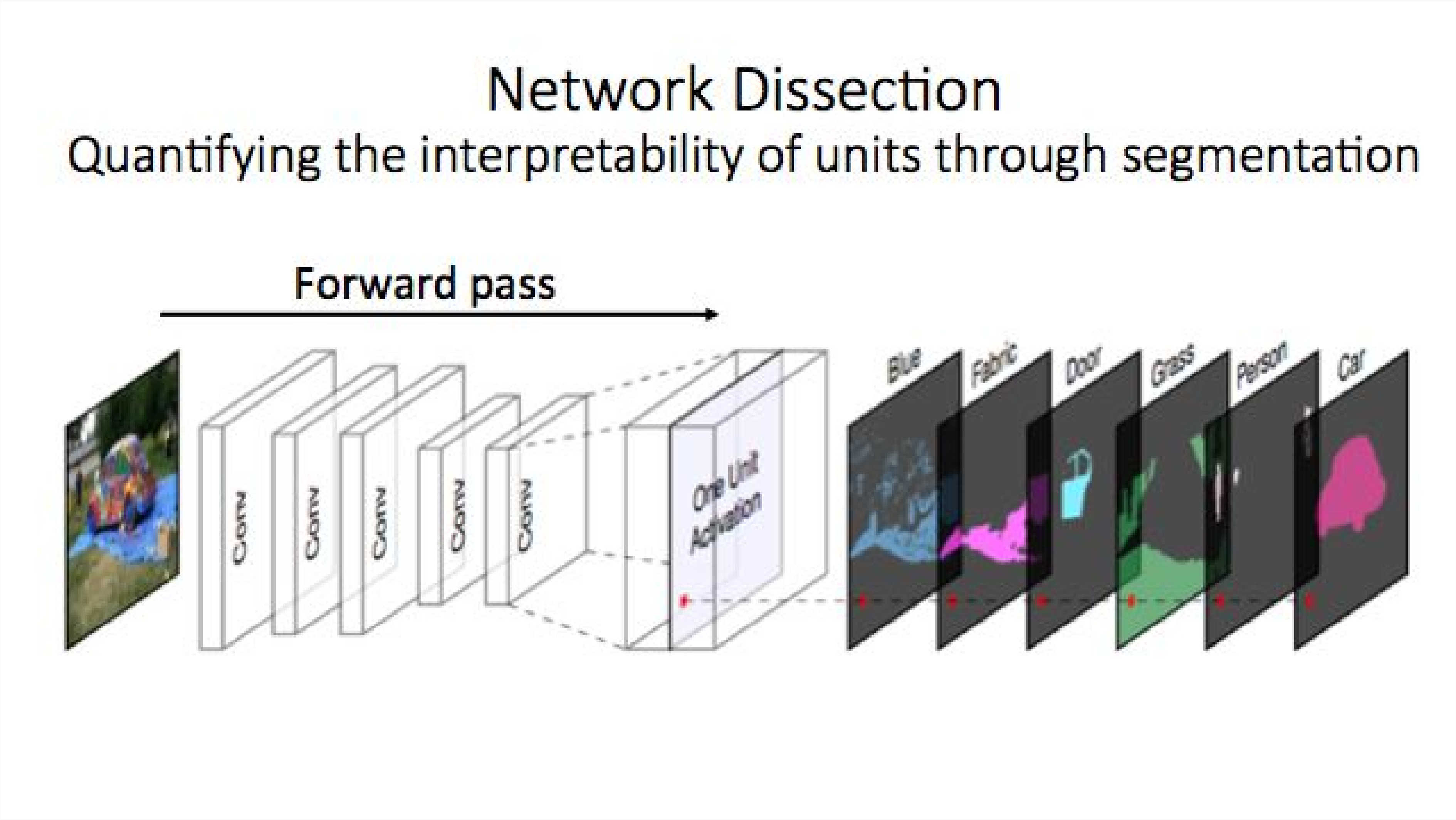
Network Dissection: Quantifying Interpretability of Deep Visual Representations

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

CS 381V

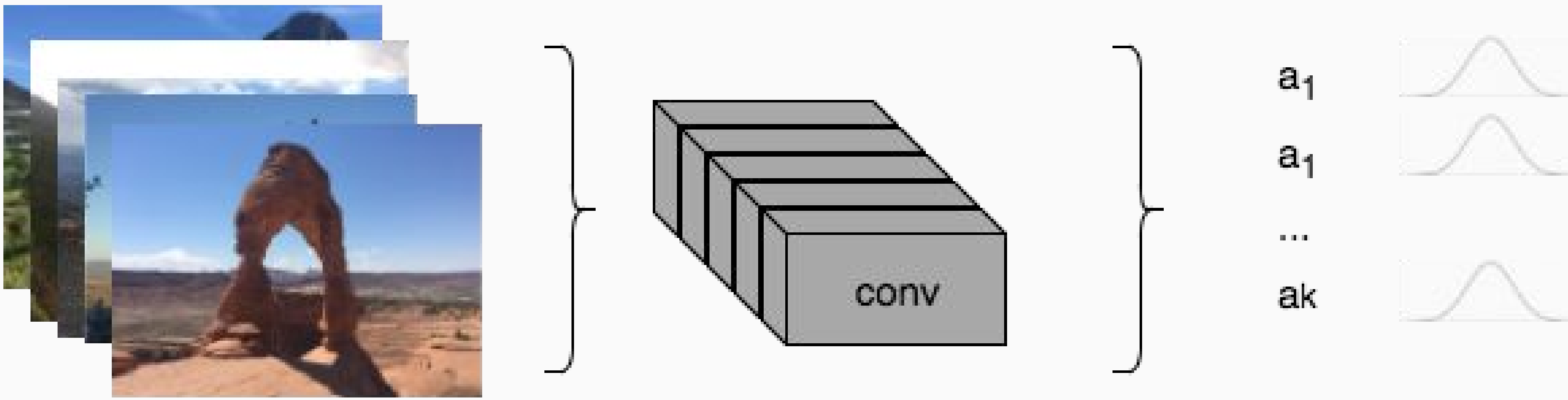
Thomas Crosley and Wonjoon Goo

Detectors



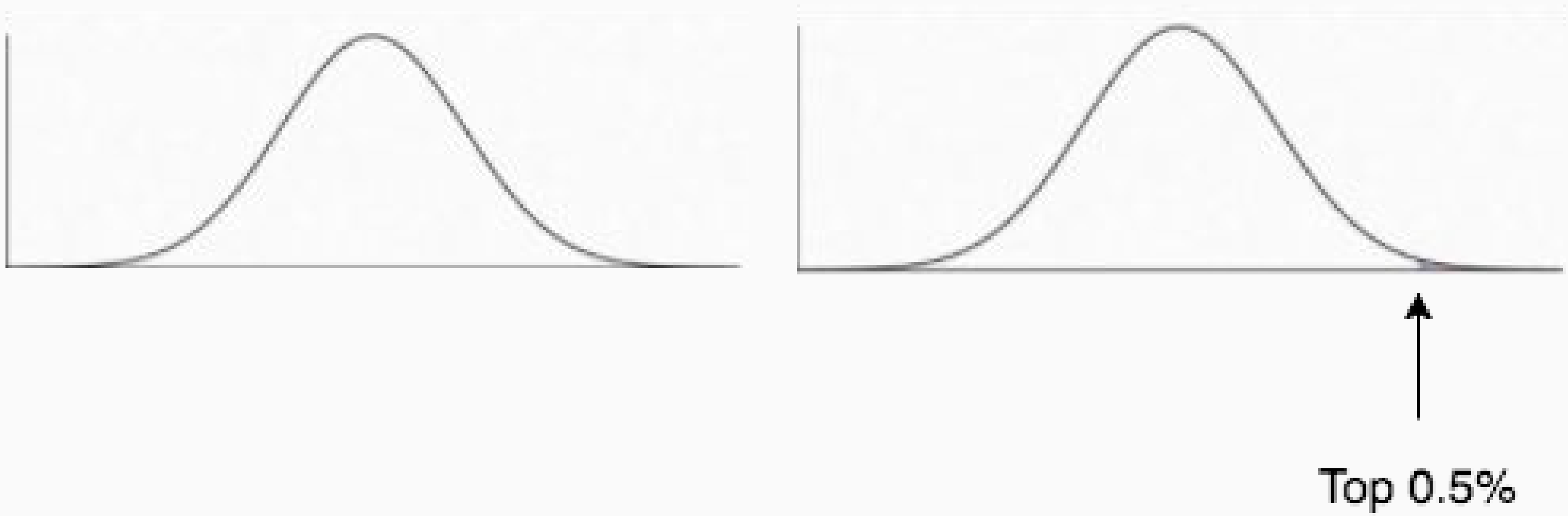
[Credit: slide from the original paper]

Unit Distributions



- Compute internal activations for entire dataset
- Gather distribution for each unit across dataset

Top Quantile



- Compute T_k such that $P(a_k > T_k) = 0.005$
- T_k is considered the top-quantile
- Detected regions at test time are those with

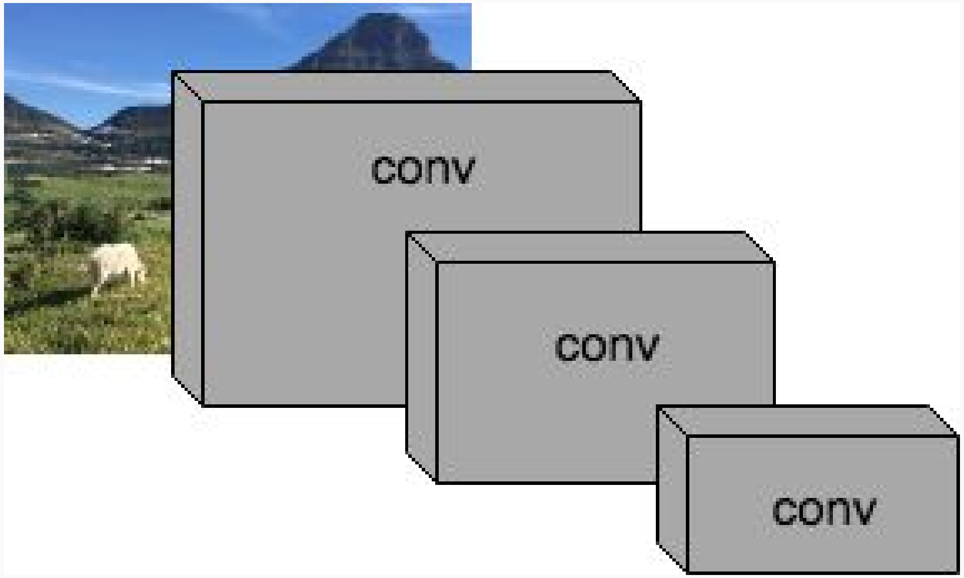
  $a_k > T_k$

Detector Concept

- Score of each unit is its IoU with the label
- Detectors are selected with IoU above a threshold
- Threshold is $U_{k,c} > 0.04$.

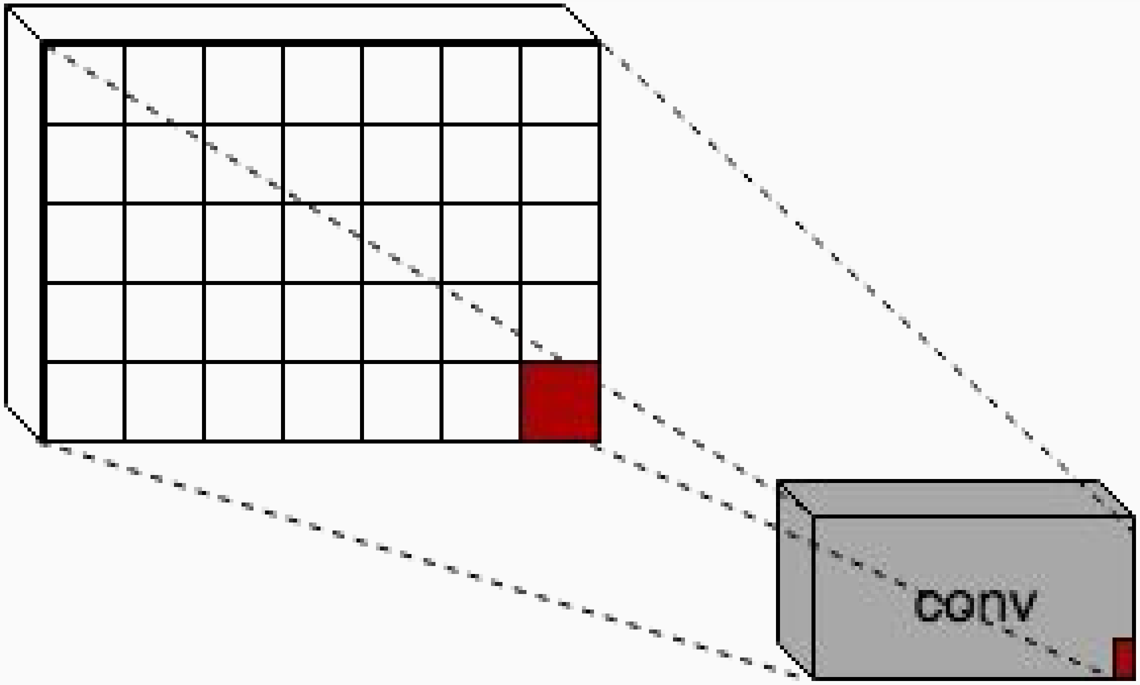
$$IoU_{k,c} = \frac{\sum |M_k(\mathbf{x}) \cap L_c(\mathbf{x})|}{\sum |M_k(\mathbf{x}) \cup L_c(\mathbf{x})|}$$

Test Data



- Compute activation map a_k for all k neurons in the network

Scaling Up



- Scale each unit's activation up to the original image size
- Call this the mask-resolution S_K
- Use bi-linear interpolation

Thresholding

99	100	100	70	68
99	100	101	67	65
98	102	100	98	60
82	97	99	97	52
75	70	72	70	45

S_K

$T_k=95$

1	1	1	0	0
1	1	1	0	0
1	1	1	1	0
0	1	1	1	0
0	0	0	0	0

M_K

- Now make the binary segmentation mask M_k
- $M_k = S_K > T_K$

Experiment: Detector Robustness

- Interest in adversarial examples
- Invariance to noise
- Composition by parts or statistics

Noisy Images

+ Unif[0, 1]

+ 5 * Unif[0, 1]

+ 10 * Unif[0, 1]

+ 100 * Unif[0, 1]

Conv3

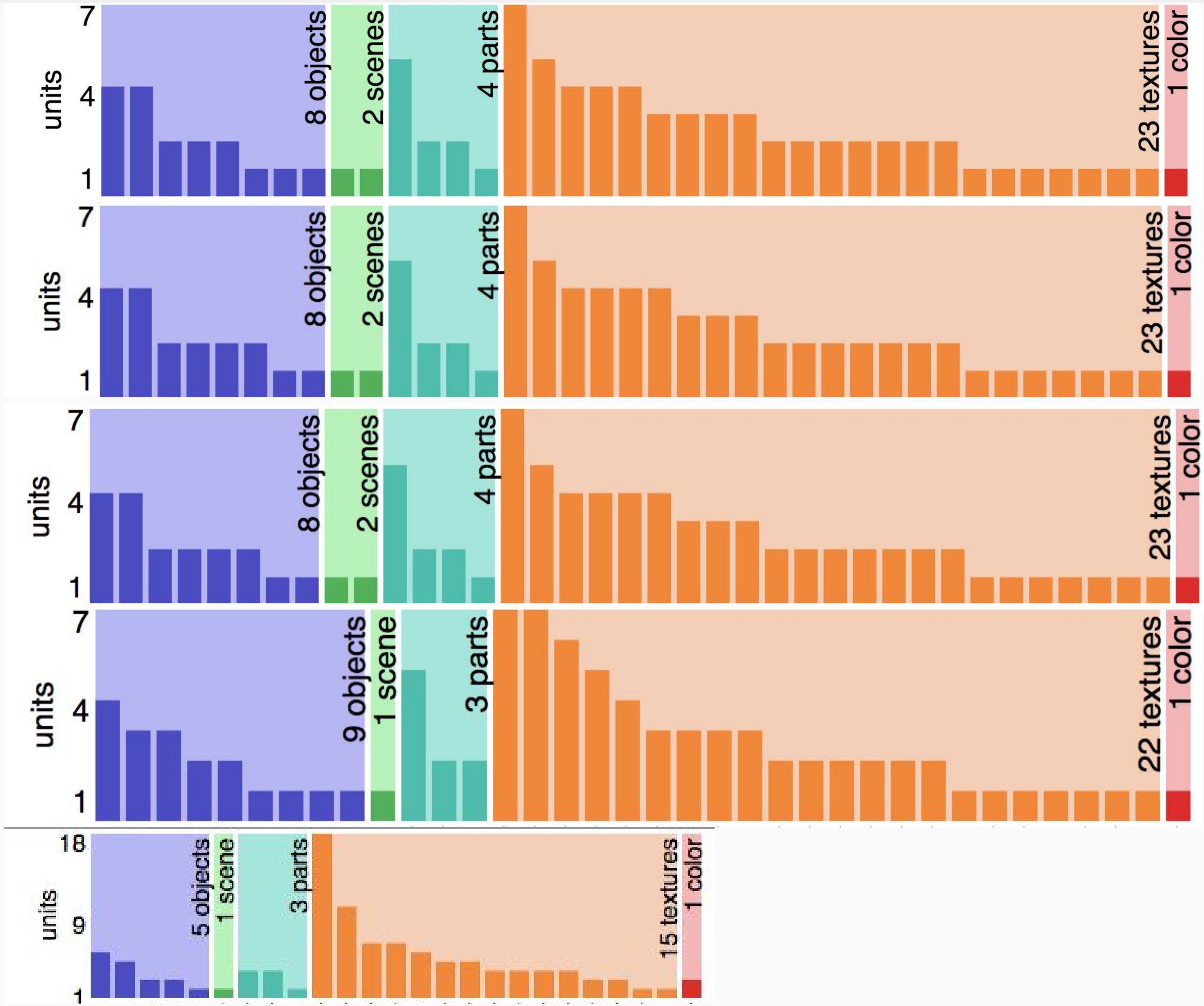
Original

+ Unif[0, 1]

+ 5 * Unif[0, 1]

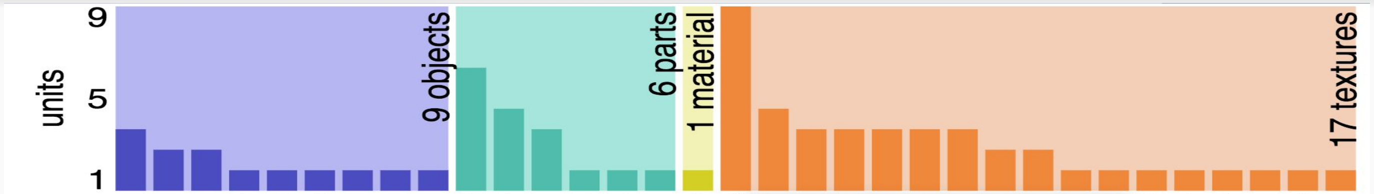
+ 10 * Unif[0, 1]

+ 100 * Unif[0, 1]

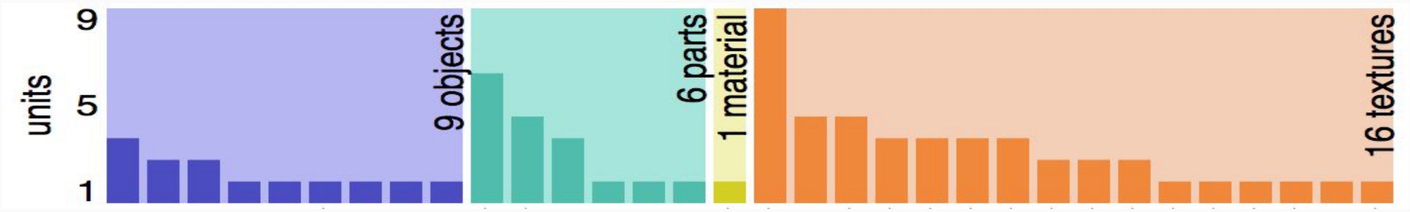


Conv4

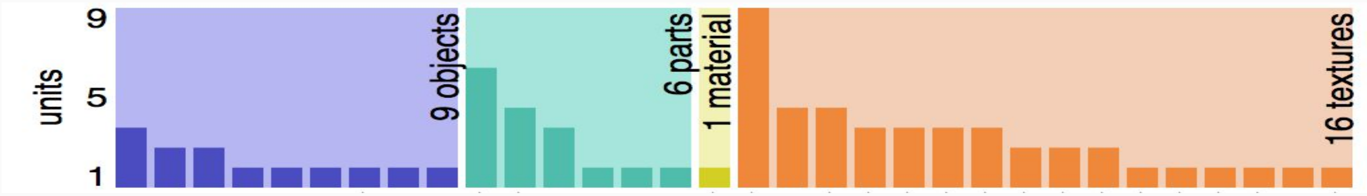
Original



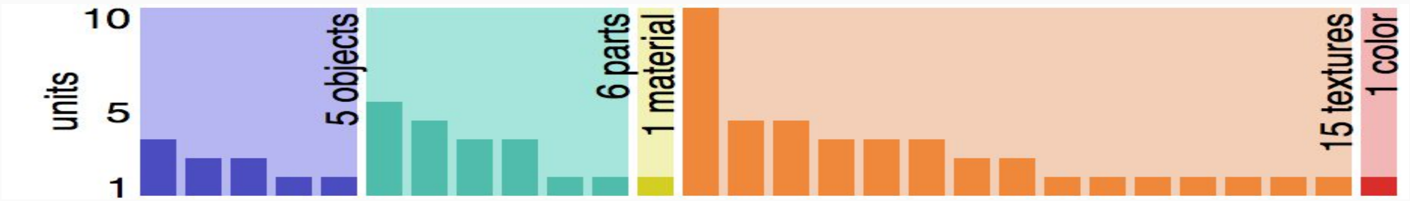
\+ Unif[0, 1]



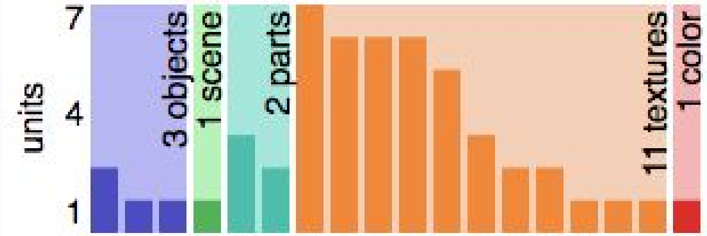
\+ 5 * Unif[0, 1]



\+ 10 * Unif[0, 1]

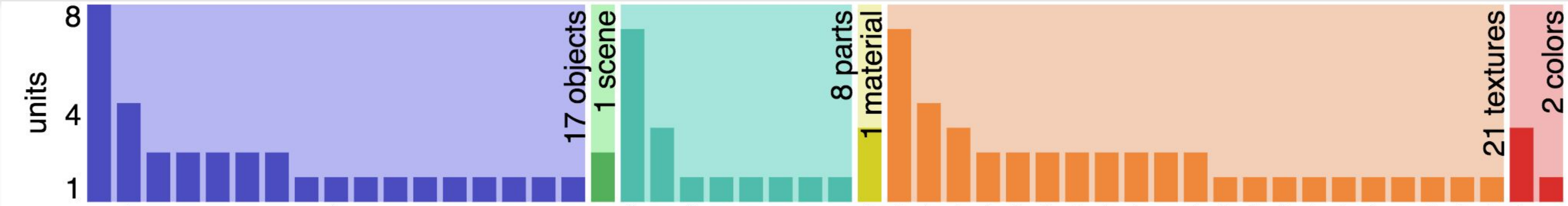


\+ 100 * Unif[0, 1]

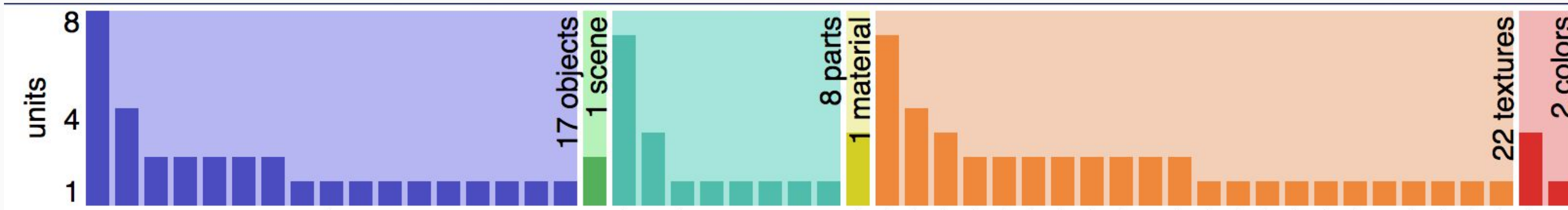


Conv5

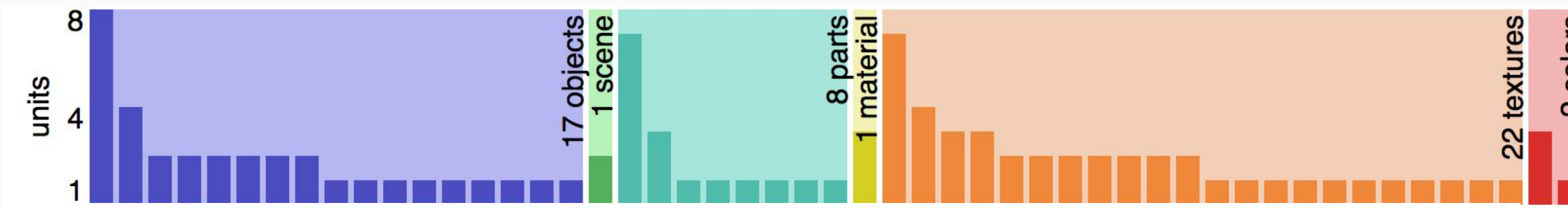
Original



\+ Unif[0, 1]

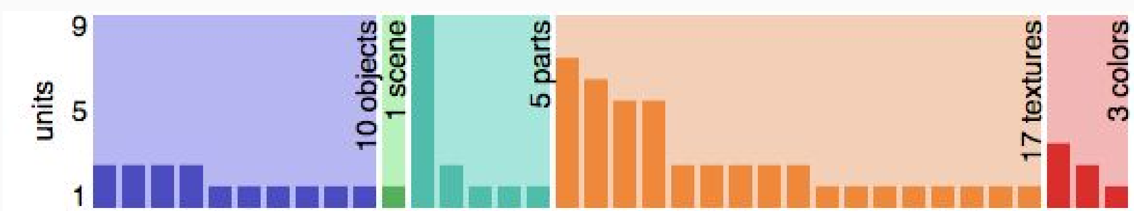


\+ 5 * Unif[0, 1]



\+ 10 * Unif[0, 1]

\+ 100 * Unif[0, 1]



Rotated Images

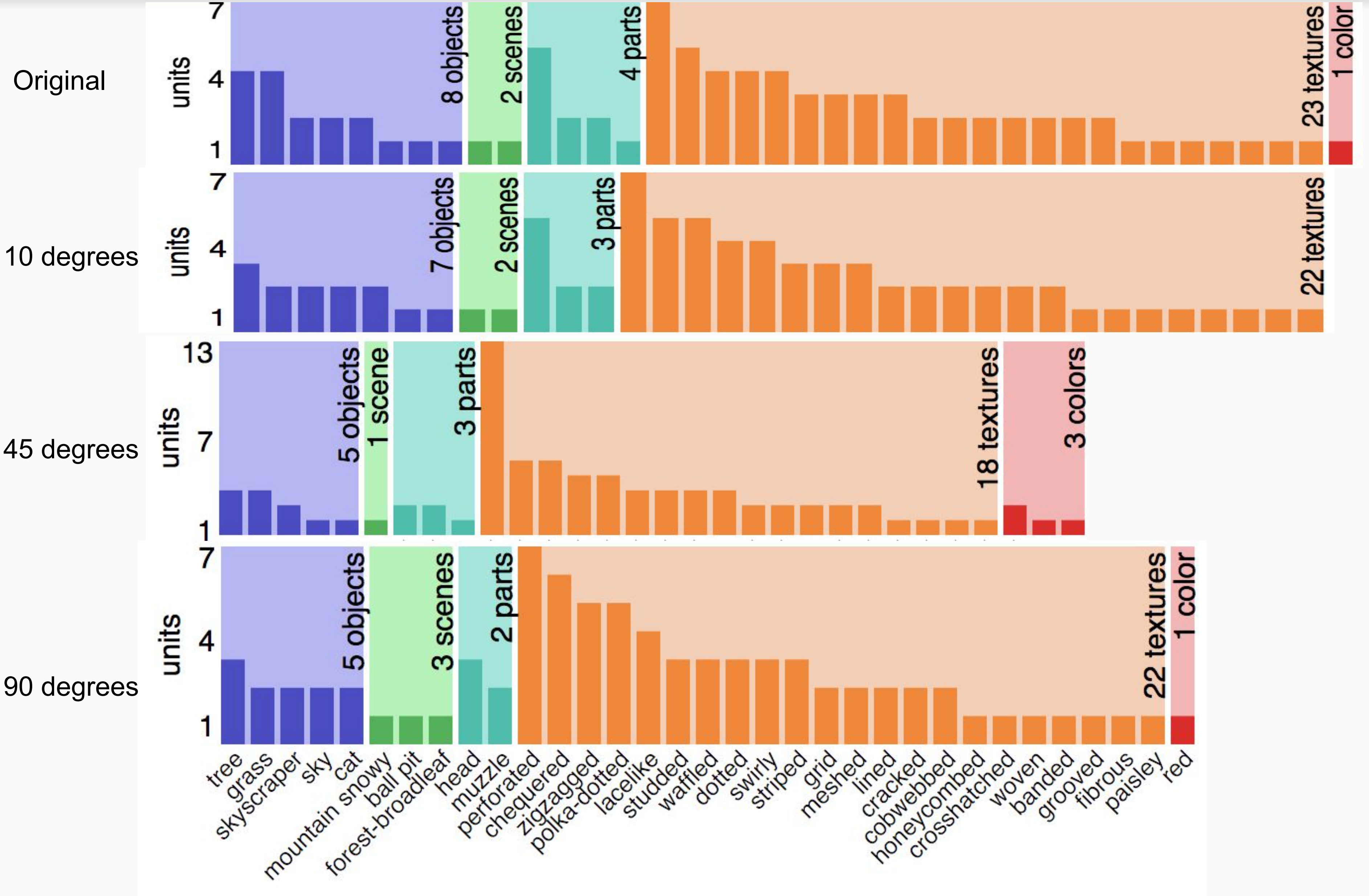
Original

10 degrees

45 degrees

90 degrees

conv3


Original
10 degrees
45 degrees
90 degrees
units
8 objects
2 scenes
4 parts
23 textures
1 color
7 objects
2 scenes
3 parts
22 textures
5 objects
1 scene
3 parts
18 textures
3 colors
5 objects
3 scenes
2 parts
22 textures
1 color
tree
grass
skyscraper
sky
cat
mountain snowy
ball pit
forest-broadleaf
head
muzzle
perforated
chequered
zigzagged
polka-dotted
lacelike
studded
waffled
dotted
swirly
striped
grid
meshed
lined
cracked
cobwebbed
honeycombed
crosshatched
woven
banded
grooved
fibrous
paisley
red

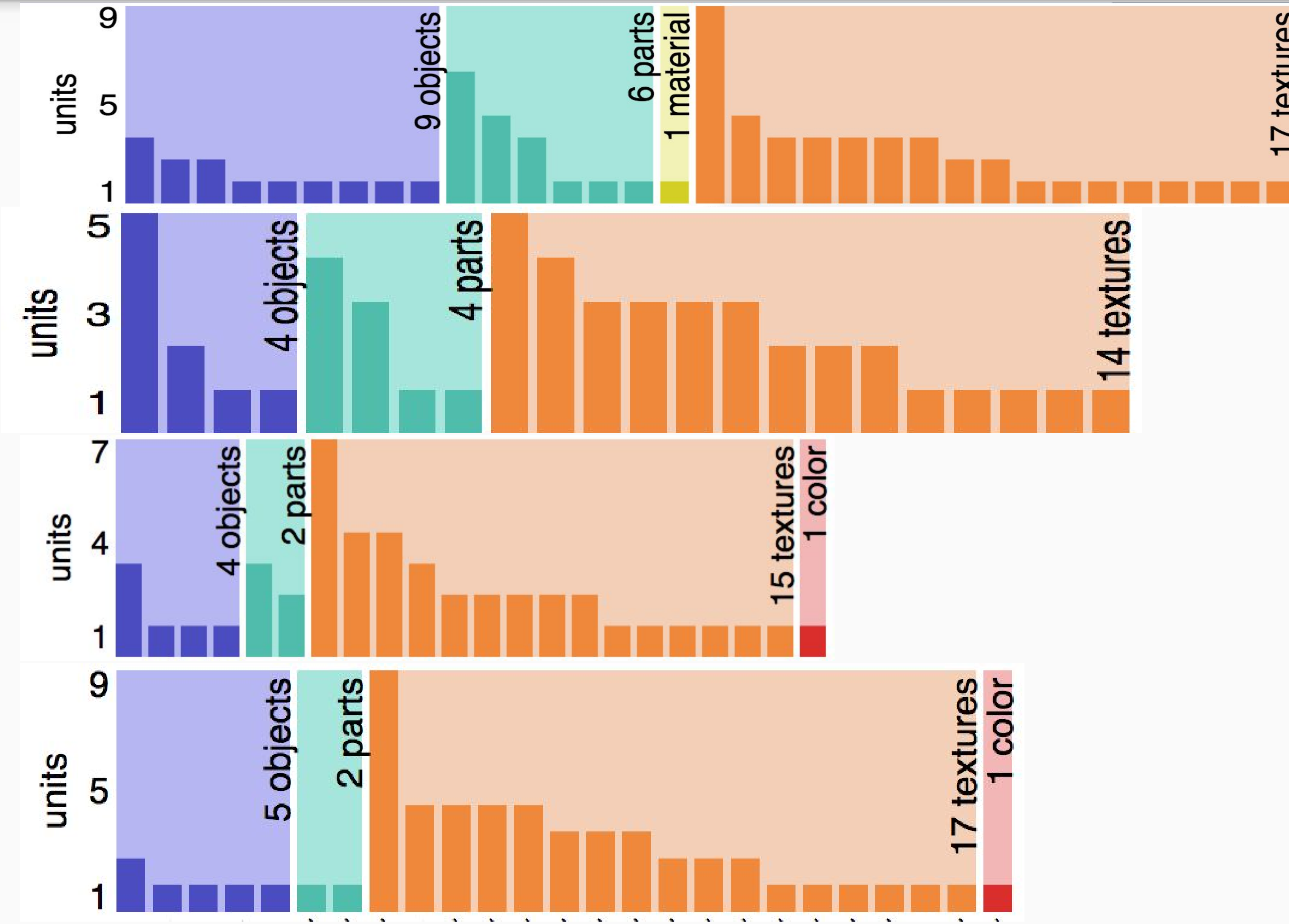
conv4

Original

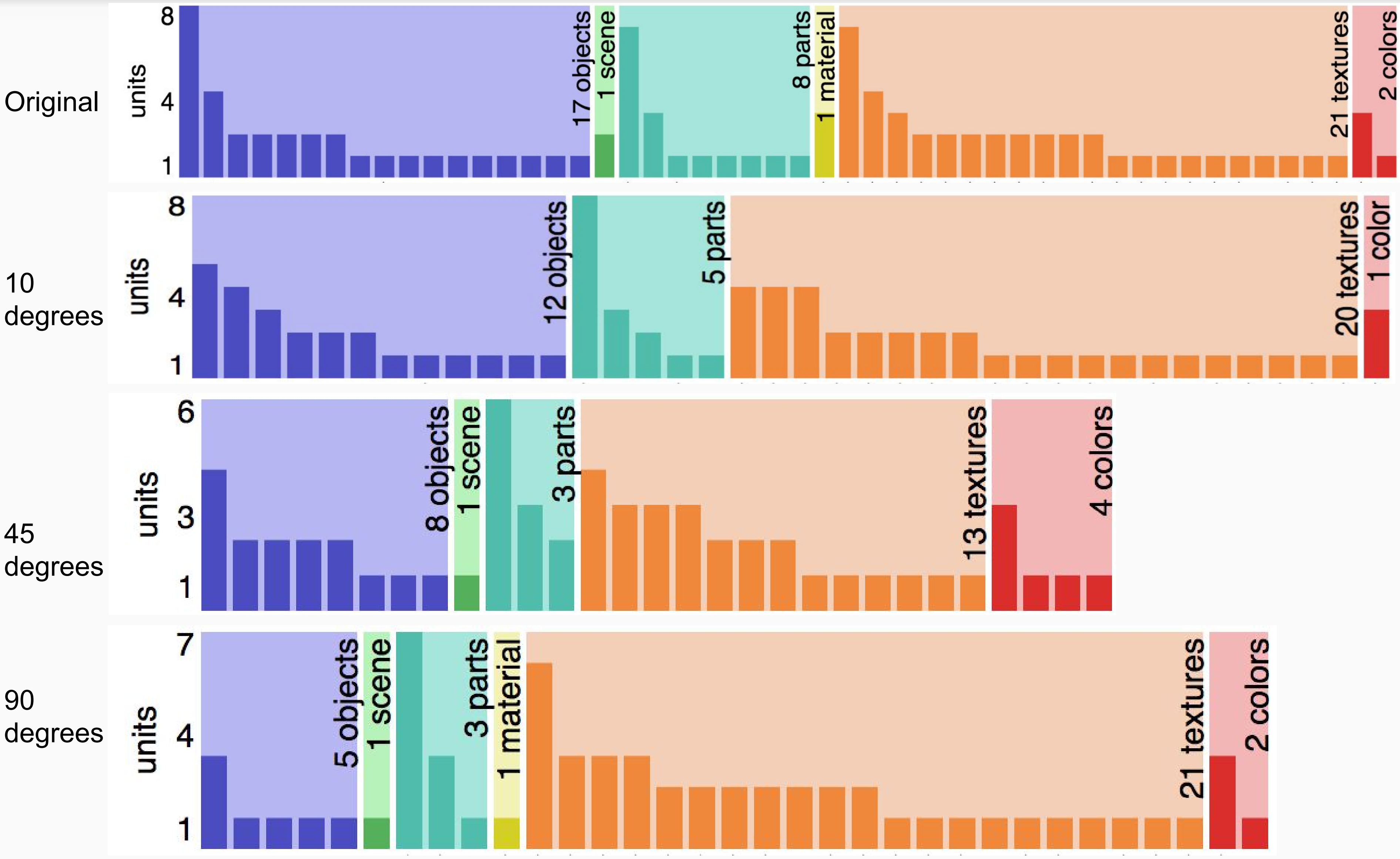
10 degrees

45 degrees

90 degrees



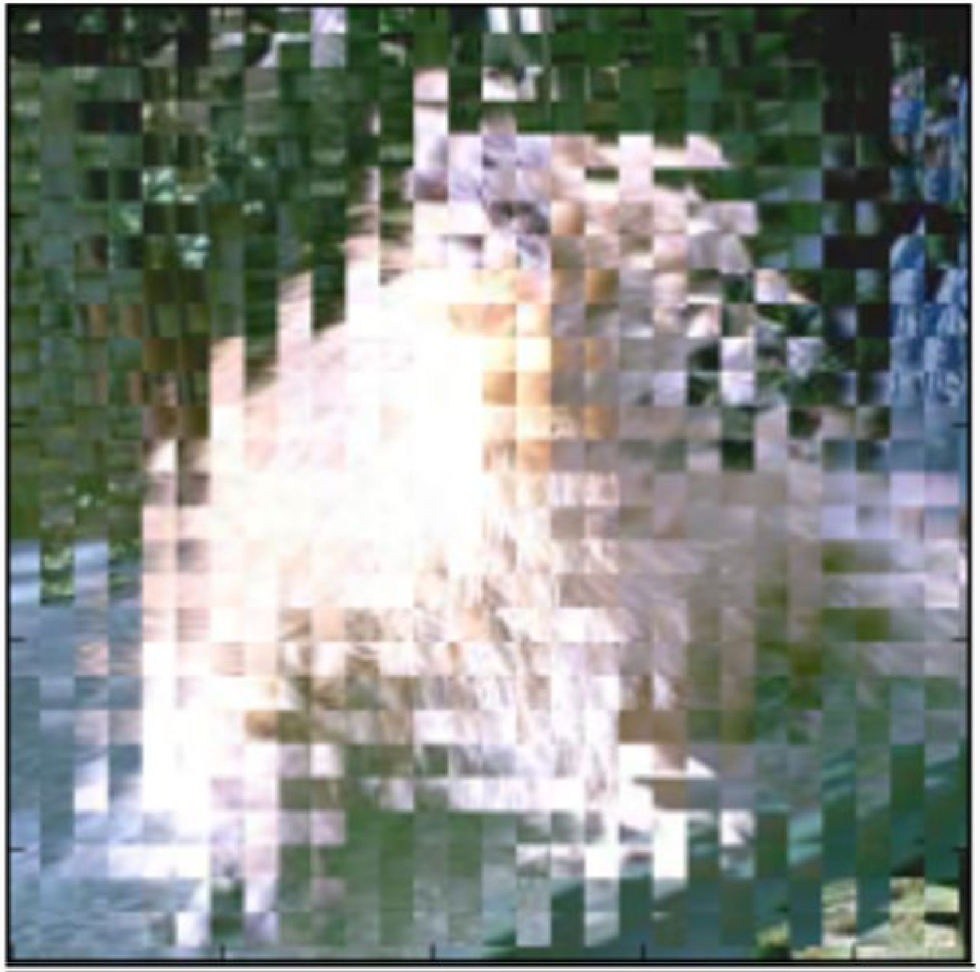
conv5


Original
units
8
4
1
17 objects
1 scene
8 parts
1 material
21 textures
2 colors
10 degrees
units
8
4
1
12 objects
5 parts
20 textures
1 color
45 degrees
units
6
3
1
8 objects
1 scene
3 parts
13 textures
4 colors
90 degrees
units
7
4
1
5 objects
1 scene
3 parts
1 material
21 textures
2 colors

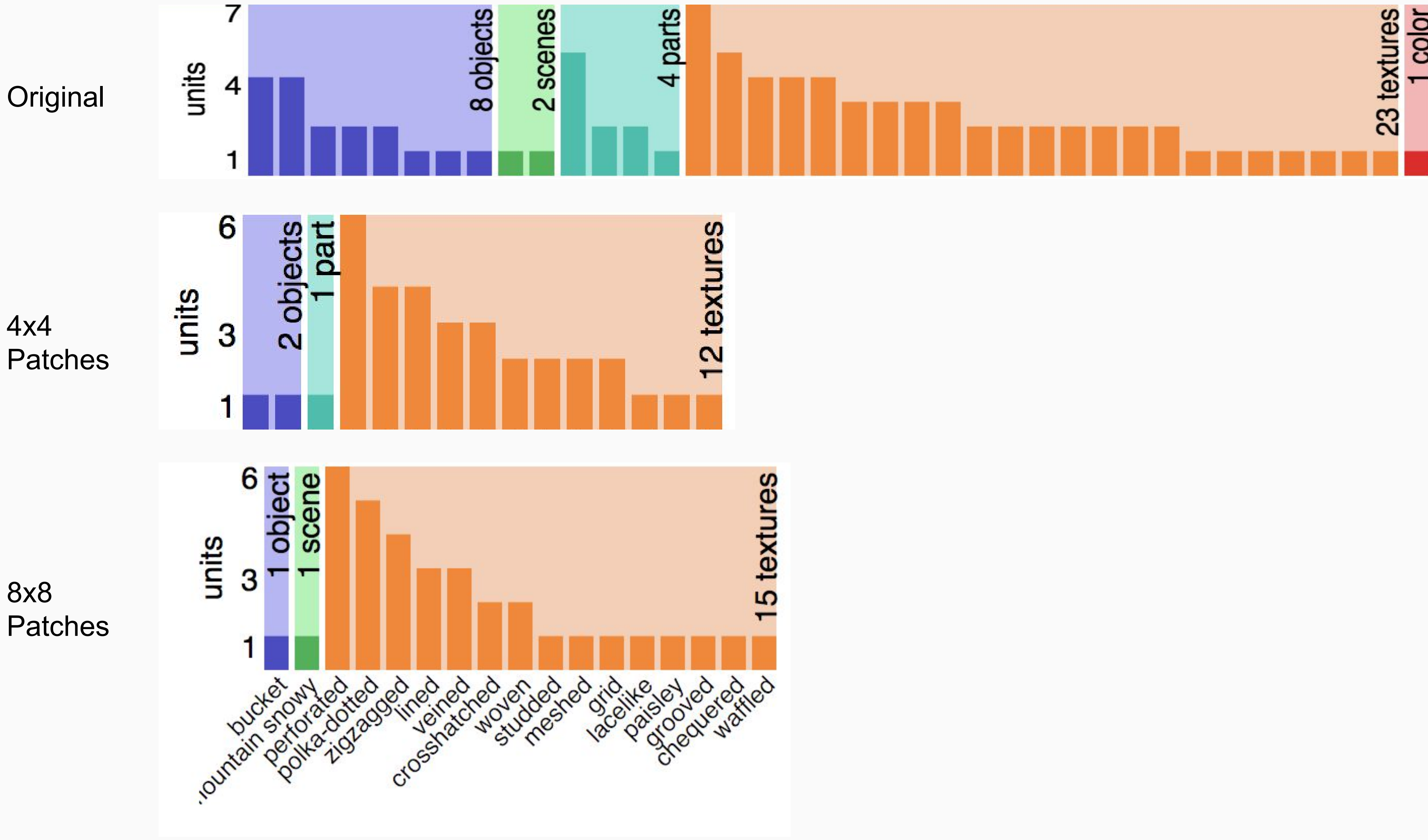
Rearranged Images

Rearranged Images

Rearranged Images



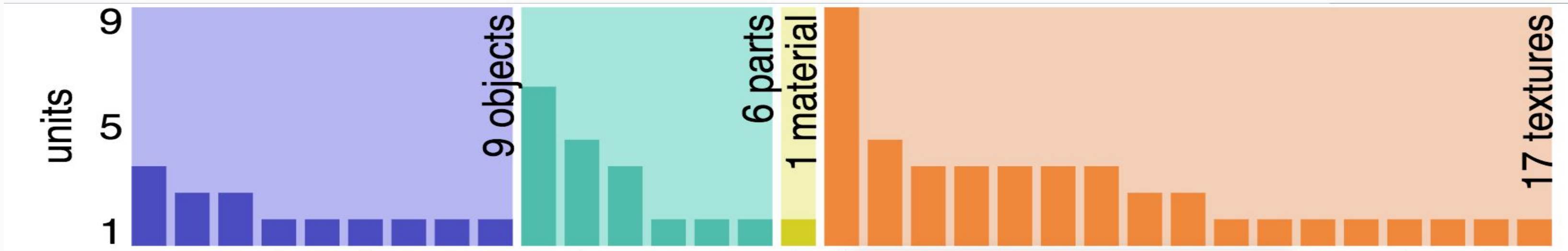
Conv3



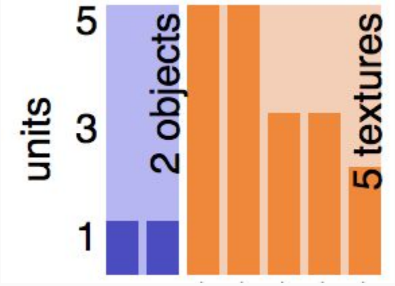
Original
4x4
Patches
8x8
Patches
units
7
4
1
8 objects
2 scenes
4 parts
23 textures
1 color
6
3
1
2 objects
1 part
12 textures
1 object
1 scene
15 textures
bucket
mountain snowy
perforated
polka-dotted
zigzagged
lined
veined
crosshatched
woven
studded
meshed
grid
lacelike
paisley
grooved
chequered
waffled

Conv4

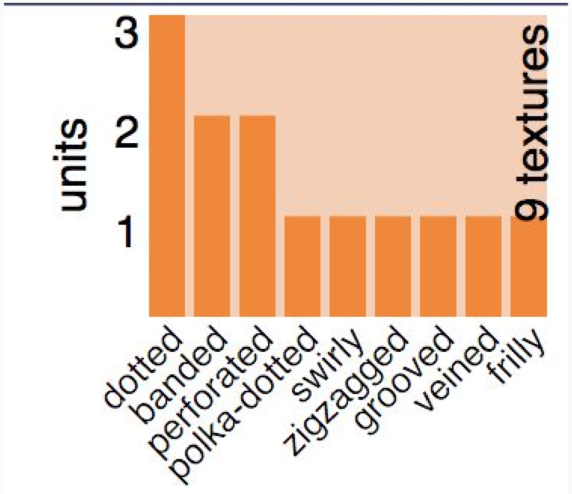
Original



4x4 Patches

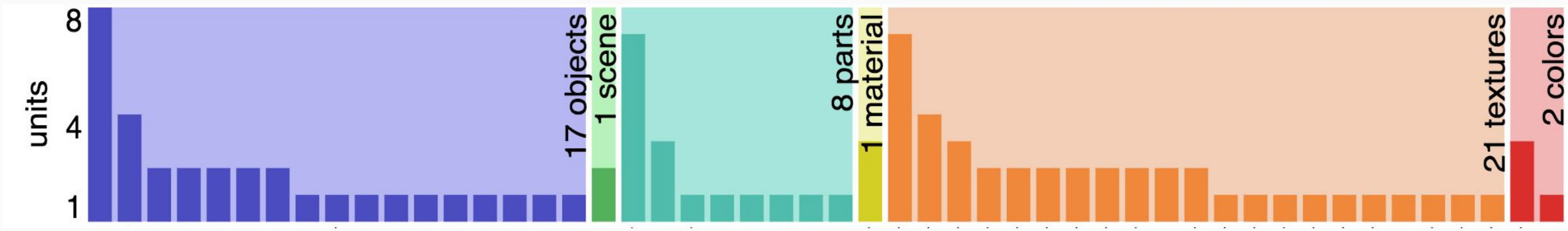


8x8 Patches

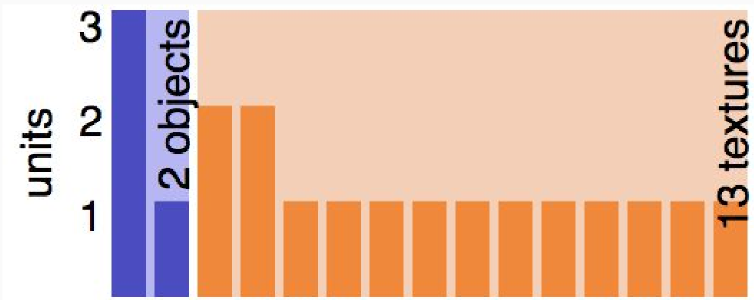


Conv5

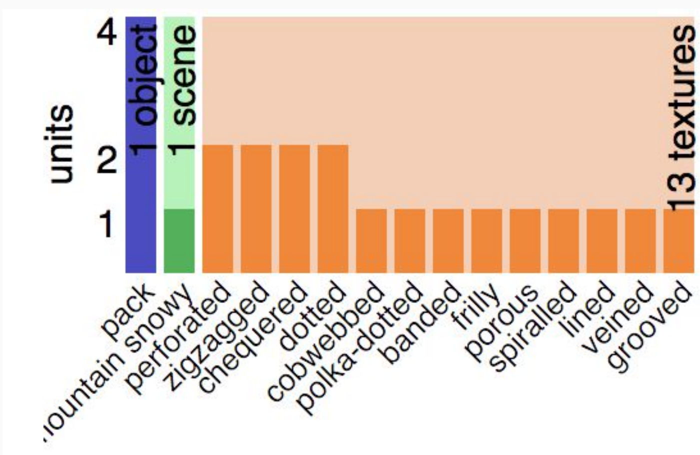
Original



4x4 Patches



8x8 Patches



Axis-Aligned Interpretability

Axis-Aligned Interpretability

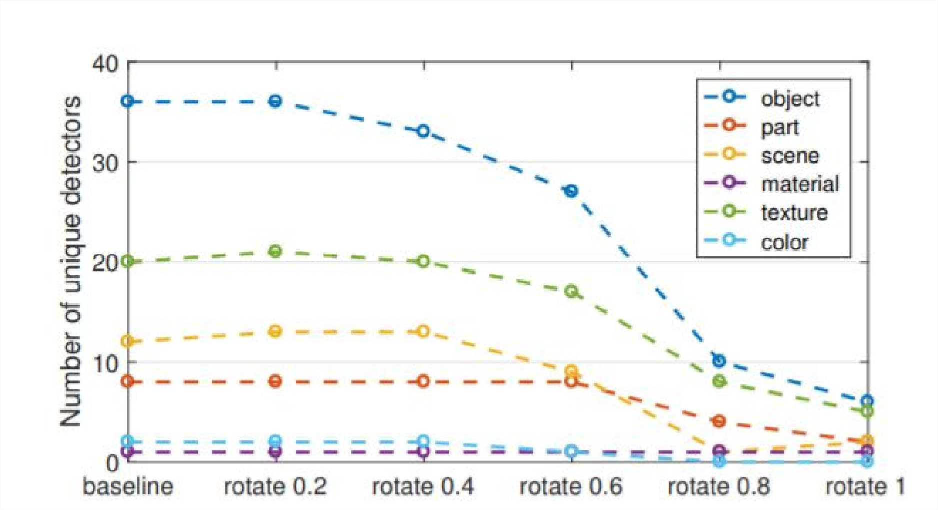
- Hypothesis 1:
  - A linear combination of high level units serves just same or better
  - No specialized interpretation for each unit

- Hypothesis 2: (the authors' argument)
  - A linear combination will degrade the interpretability
  - Each unit serves for unique concept

How similar is the way CNN learns to human?

Axis-Aligned Interpretability Result from the Authors

Figure: from the paper

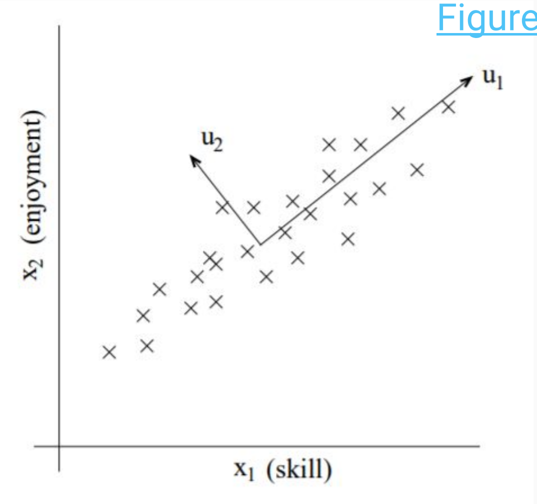


- It seems valid argument, but is it the best way to show?
- Problems
    - It depends on a rotation matrix used for test
    - A 90 degree rotation between two axis, does not affect the number of unique detectors
    - The test should be done multiple times and report the means and stds.

Experiment: Axis-Aligned Interpretability

Is it really axis aligned?

Figure: From Andrew Ng's lecture note on PCA



- Principle Component Analysis (PCA)
  - Find orthonormal vectors explaining samples the most
  - The projections to the vector u_1 have higher variance

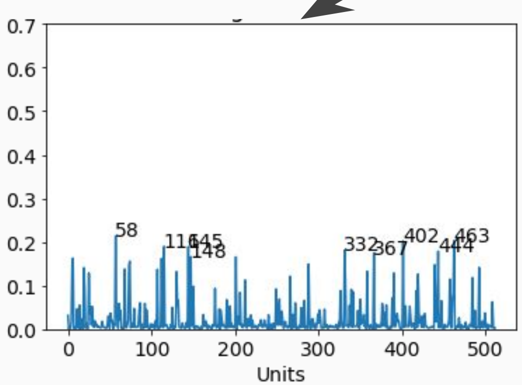
❖ Argument: a unit itself can explain a concept
  - ➢ Projections to unit vectors should have higher variance
  - ➢ *Principal axis (Loading)* from PCA should be similar to one of the unit vectors

Our method

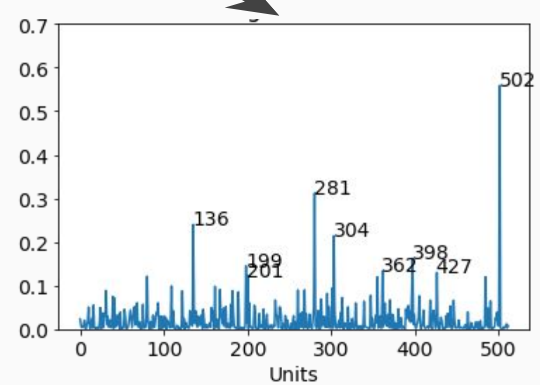
1. Calculate the mean and std. of each unit activation
2. Grab activations for a specific concept
3. Subtract mean and std from activations
4. Perform SVD
5. Print Loading

Hypothesis 1

Hypothesis 2



The concept is interpreted with the combination of elementary basis



The concept can be interpreted with an elementary basis (eg. e_502 := (0,...,0,1,0,...,0))

(Supplementary) PCA and Singular Value Decomposition (SVD)

From Cheng Li, Bingyu Wang Notes

- Optimize target:

$$\max_{u:\|u\|=1} \frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u)^2$$

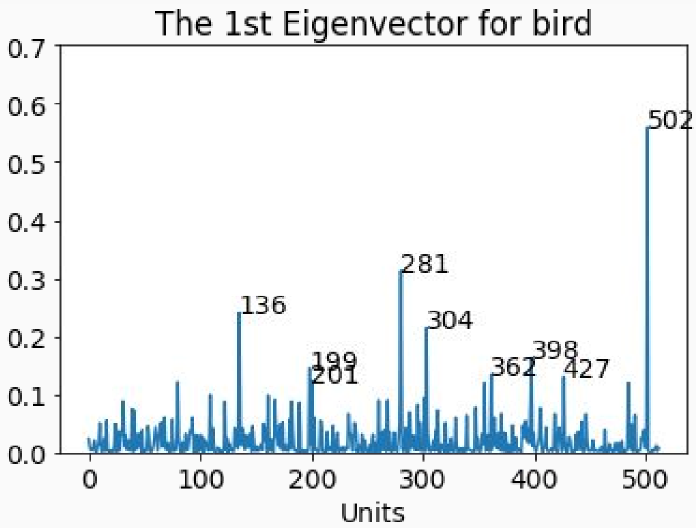
$$\begin{aligned}
&\Longrightarrow \max_{u:\|u\|=1} \frac{1}{m}\sum_{i=1}^{m}(x^{(i)T}u)^T(x^{(i)T}u) \\
&\Longrightarrow \max_{u:\|u\|=1} \frac{1}{m}\sum_{i=1}^{m}(u^T x^{(i)})(x^{(i)T}u) \\
&\Longrightarrow \max_{u:\|u\|=1} \frac{1}{m}\sum_{i=1}^{m}u^T x^{(i)}x^{(i)T}u \\
&\Longrightarrow \max_{u:\|u\|=1} u^T(\frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)T})u \\
&\Longrightarrow \max_{u:\|u\|=1} u^T\Sigma u, \quad where \quad \Sigma = \frac{1}{m}\sum_{i=1}^{m}x^{(i)}x^{(i)T}
\end{aligned}$$

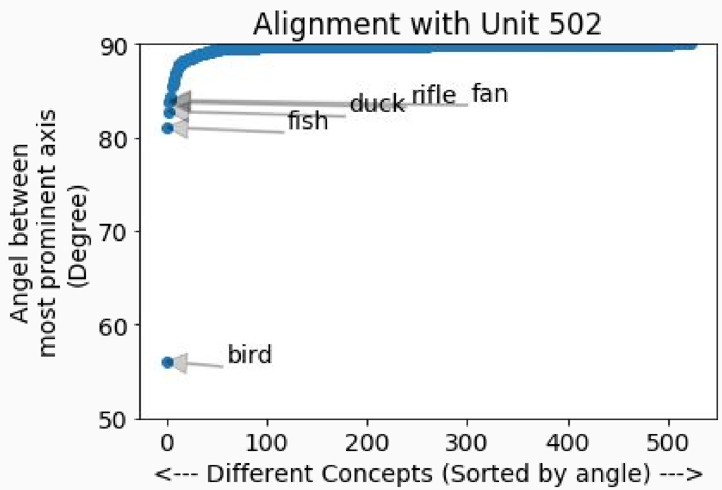
- With Lagrange multiplier:

$$\begin{aligned}
&\Longrightarrow \max u^T\Sigma u \qquad subject \quad u^Tu = 1(Because\ \|u\| = 1) \\
&\Longrightarrow \mathcal{L}(u,\lambda) = u^T\Sigma u + \lambda(u^Tu - 1) \\
&\Longrightarrow \frac{\partial\mathcal{L}(u,\lambda)}{\partial u} = \frac{\partial(u^T\Sigma u + \lambda u^Tu)}{\partial u} = \frac{\partial(u^T\Sigma u)}{\partial u} + \frac{\partial(\lambda u^Tu)}{\partial u} \\
&= 2\Sigma u + 2\lambda u \overset{Set}{=} 0
\end{aligned}$$

- The eigenvector for the highest eigenvalue becomes principal axis (loading)

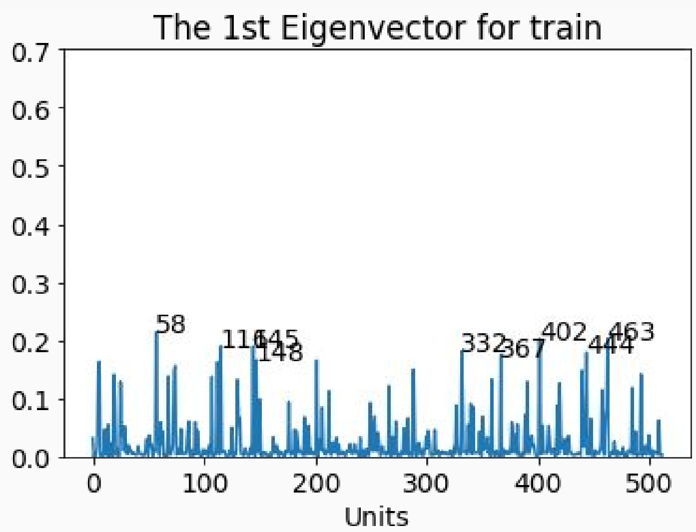
PCA Results - Activations for Bird Concept





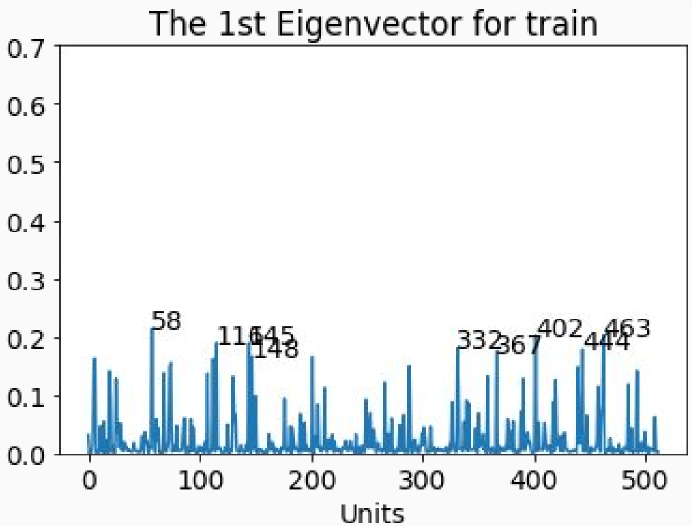
- Unit 502 stands high; concept bird is aligned to the unit
- Does Unit 502 only serve for concept Bird?
  - Yes
  - It does not stand for other concepts except bird
- Support Hypothesis 2

PCA Results - Activations for Train Concept



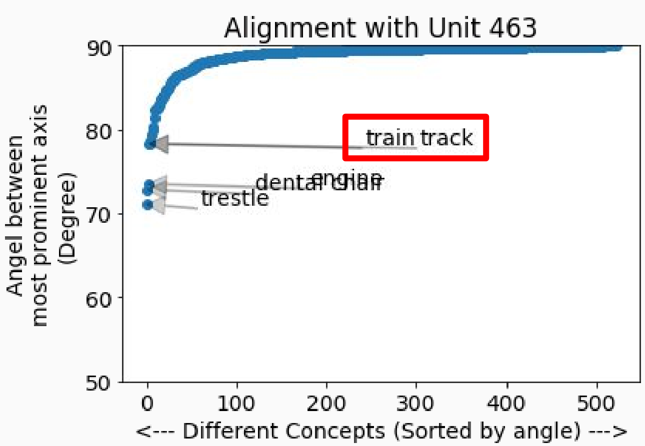
- No units stands out for concept train
  - Linear combination of them have better interpretability
  - Support Hypothesis 1

PCA Results - Activations for Train Concept

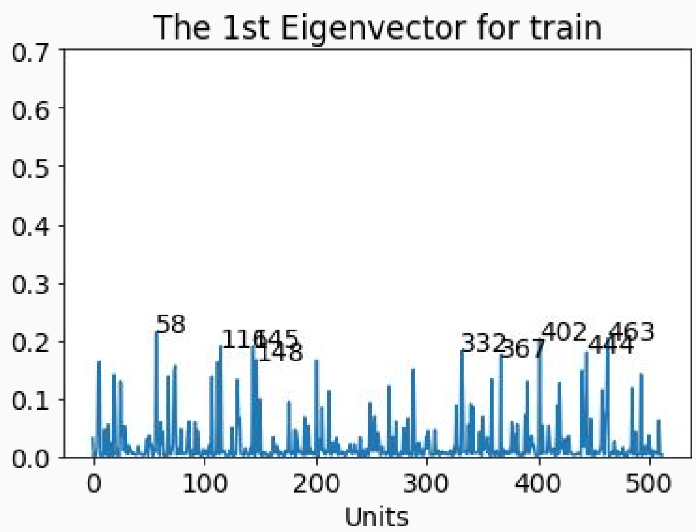


- No units stands out for concept train
  - Linear combination of them have interpretability

Some objects with circle and trestle?

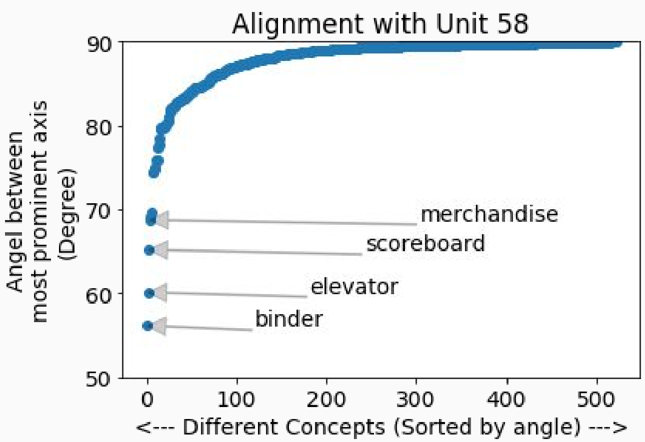


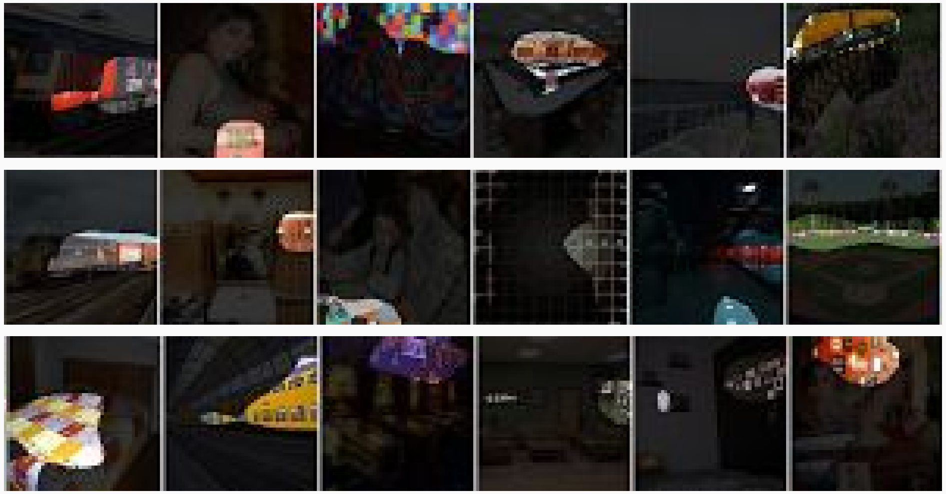
PCA Results - Activations for Train Concept



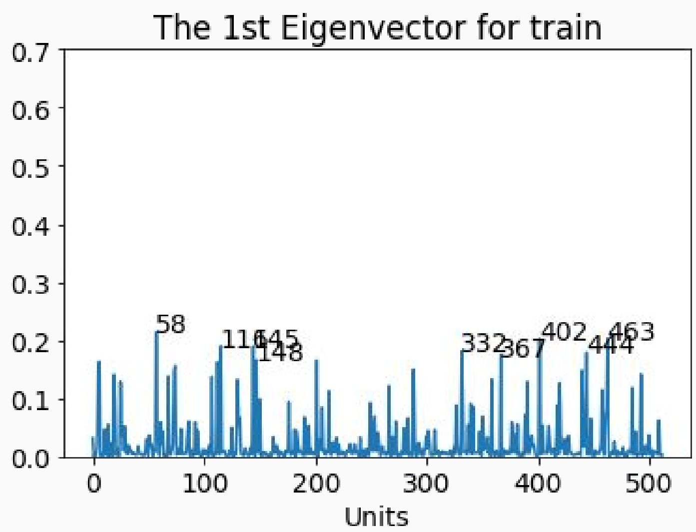
- No units stands out for concept train
    - Linear combination of them have interpretability

The sequence of square boxes?

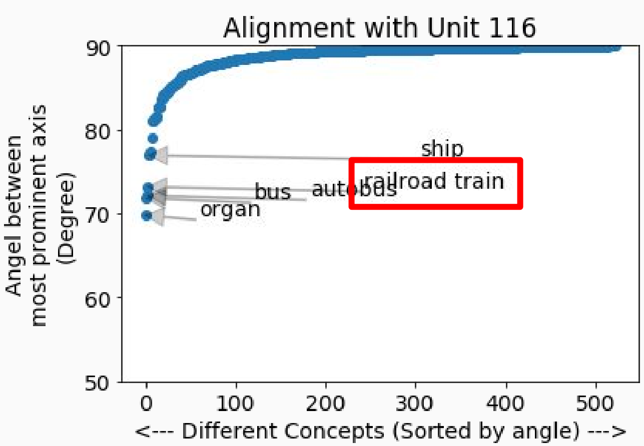




PCA Results - Activations for Train Concept



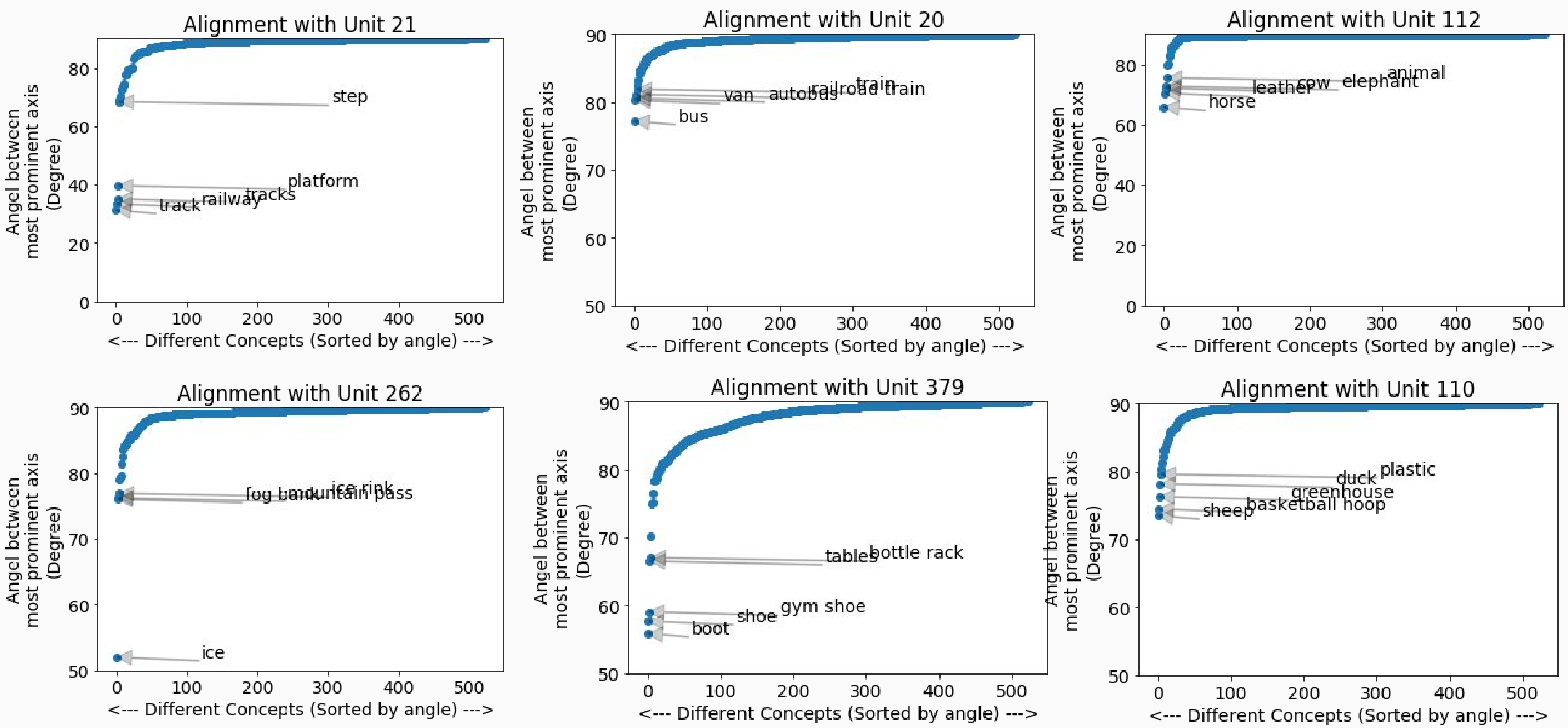
- No units stands out for concept train
  - Linear combination of them have interpretability



Dog face!

Conclusion...?

- Actually, it seems mixed!
- CNN learns some human concepts naturally, but not always
    - It might highly correlated with the label we give



Other Thoughts

- What if we regularize the network to encourage its interpretability?
  - Taxonomy-Regularized Semantic Deep Convolutional Neural Networks, Wonjoon Goo, Juyong Kim, Gunhee Kim, and Sung Ju Hwang, ECCV 2016

Thanks!
Any questions?