Systematically Figure Out the Semantic of Components in Neural Networks

Network Dissection method for investigating NN units

David Bau et al. Understanding the role of individual units in a deep neural network. Proceedings of the National Academy of Sciences Sep 2020, 201907375; DOI: 10.1073/pnas.1907375117

Tianqin Li

Oct 24, 2020

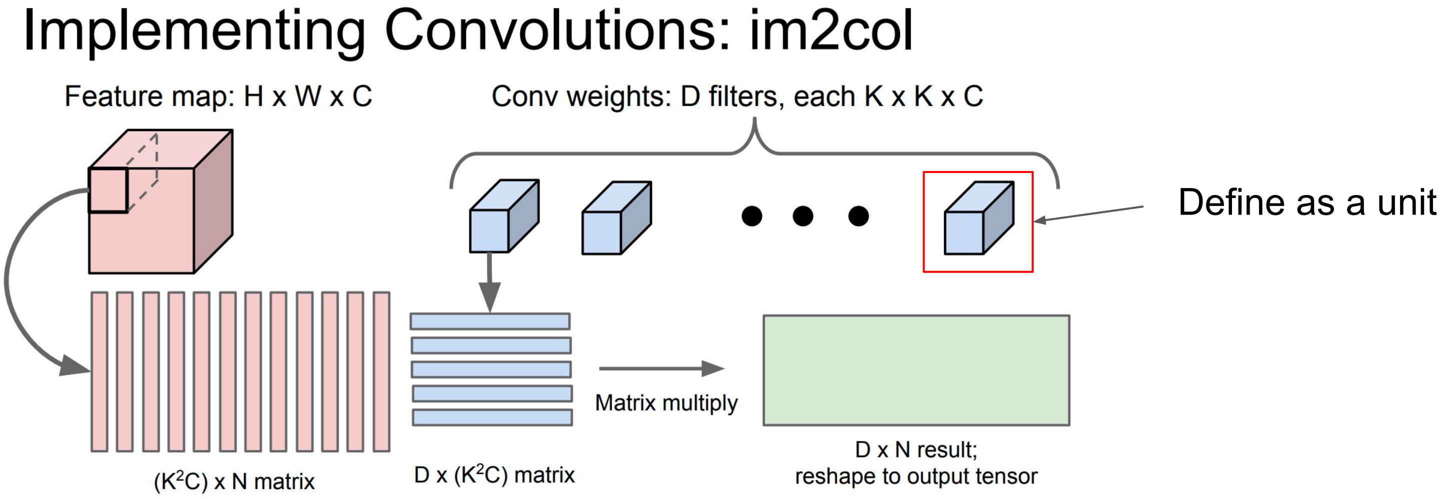
Overview

- *Network Dissection* method is developed by David Bau and Bolei Zhou at MIT Antonio Torralba's lab.
- Using image segmentation technique to investigate the causal connection between filters in CNN and human understandable visual concepts (like trees).
- Comparing to visual concept, it use a trained segmentation network to automatically find out which CNN filters are responsible for certain semantic concepts. -- wonder if one can use segmentation net to label VCs

Visual Concepts from Alan Yuille group

Network Dissection

- Define units as a filter in CNN

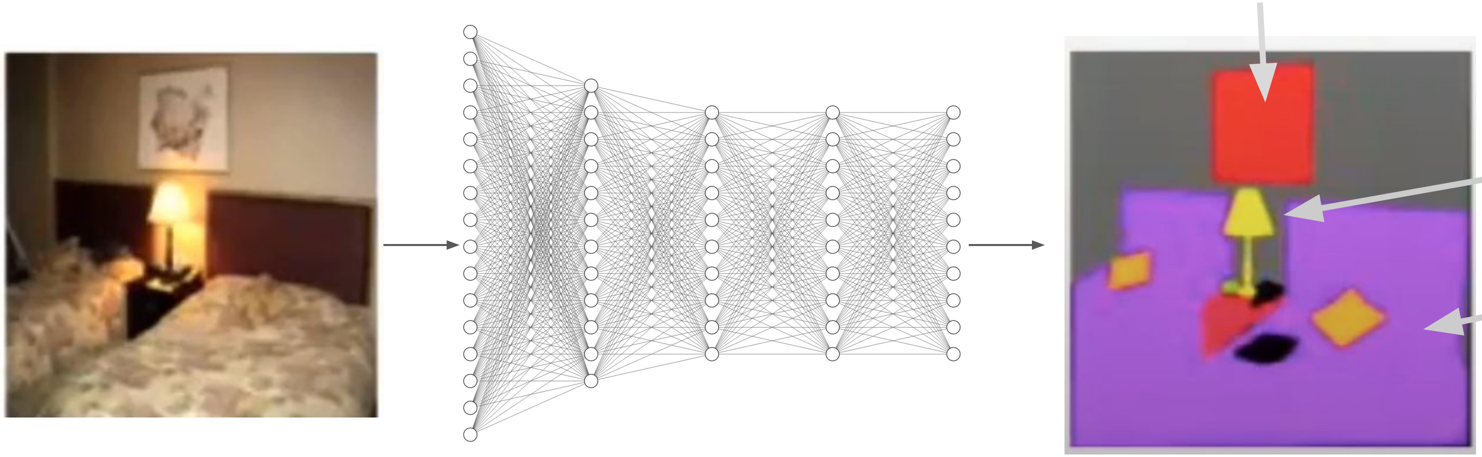


Define as a unit

http://cs231n.stanford.edu/slides/2016/winter1516_lecture11.pdf

Network Dissection

- Concepts labeled by a trained image segmentation labeling network



Picture (concept)

Lamp
(concept)

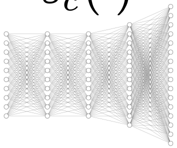
Bed
(concept)

Trained image segmentation network
MobileNetv2, ResNet, etc.

Network Dissection

- Quantifying how each individual filter (*u*) influence the detection/construction of the concepts (denote as *c*)

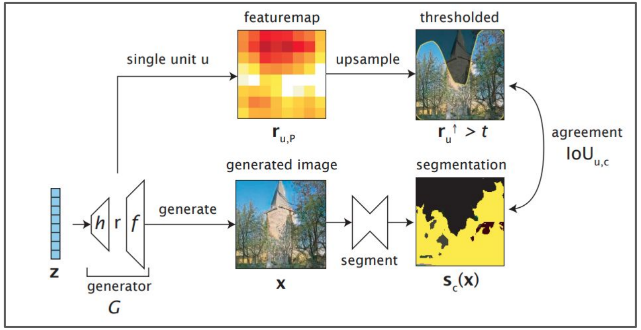
$s_c(\cdot)$



u

Activation of u, for image x at location p (scalar)

$$s_c(x, p) \in \{0, 1\} \longleftarrow x \longrightarrow a_u(x, p)$$



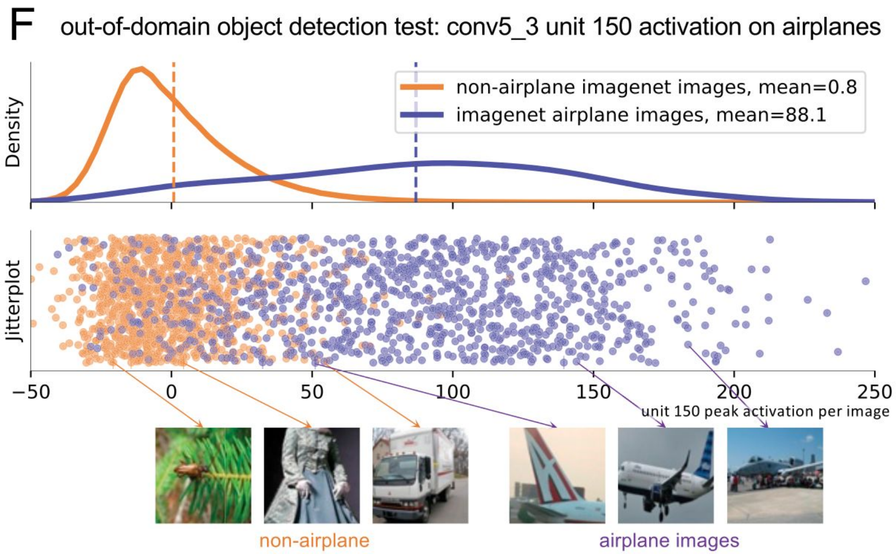
How well are the real segmentation matched with activation of the unit?

$$\mathrm{IoU}_{u,c} = \frac{\mathbb{P}_{x,p}[s_c(x,p) \wedge (a_u(x,p) > t_u)]}{\mathbb{P}_{x,p}[s_c(x,p) \vee (a_u(x,p) > t_u)]}.$$

$$t_u(threshold) \quad \max_t \mathbb{P}_{x,p}[a_u(x,p) > t] \geq 0.01.$$

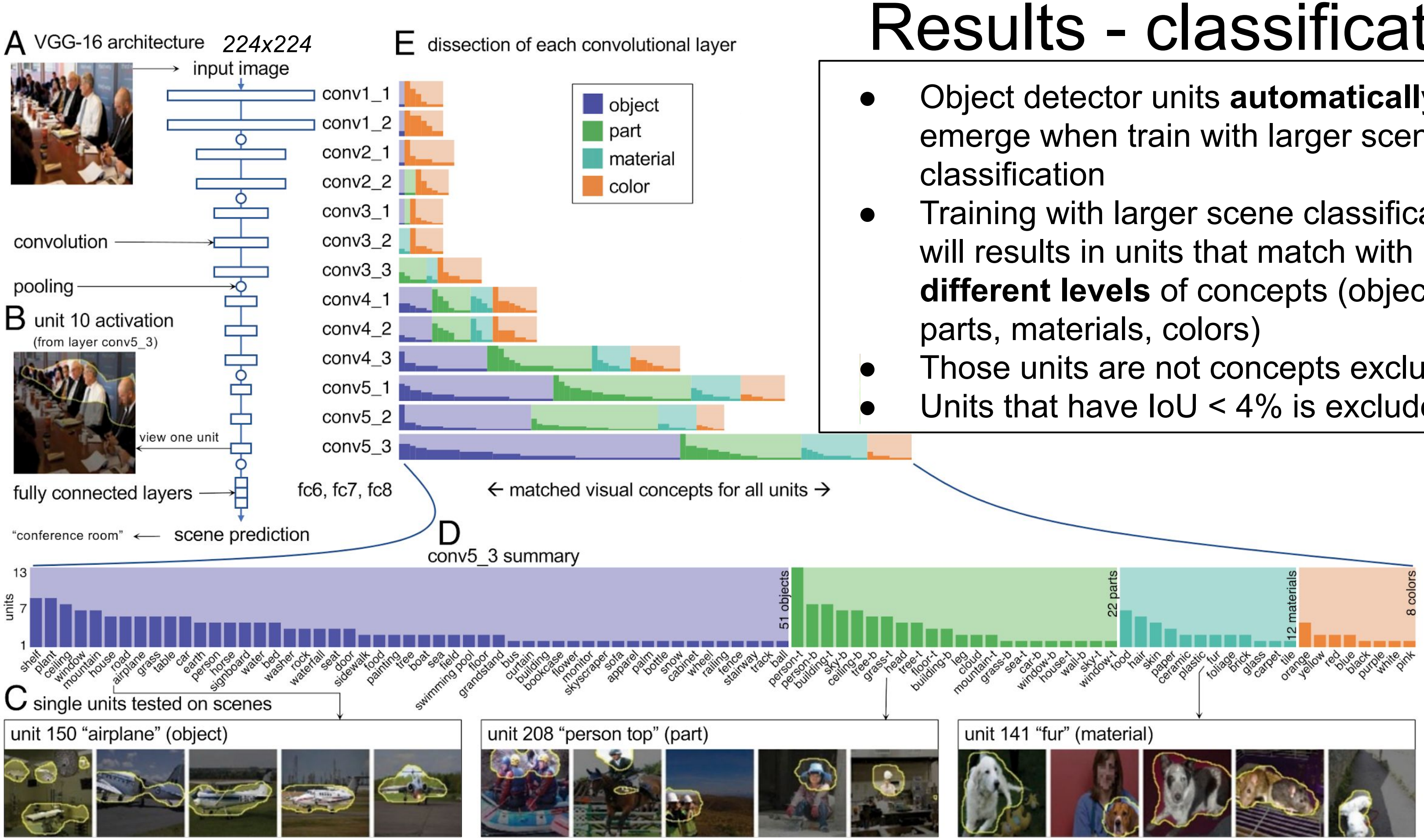
Example of one unit (u) to one concept (c)

- u is unit 150 in the last cnn layer of VGG-16 (conv5_3)
- c is airplane
- Unit 150 (u) prefer airplane concept (c)



Results - classification

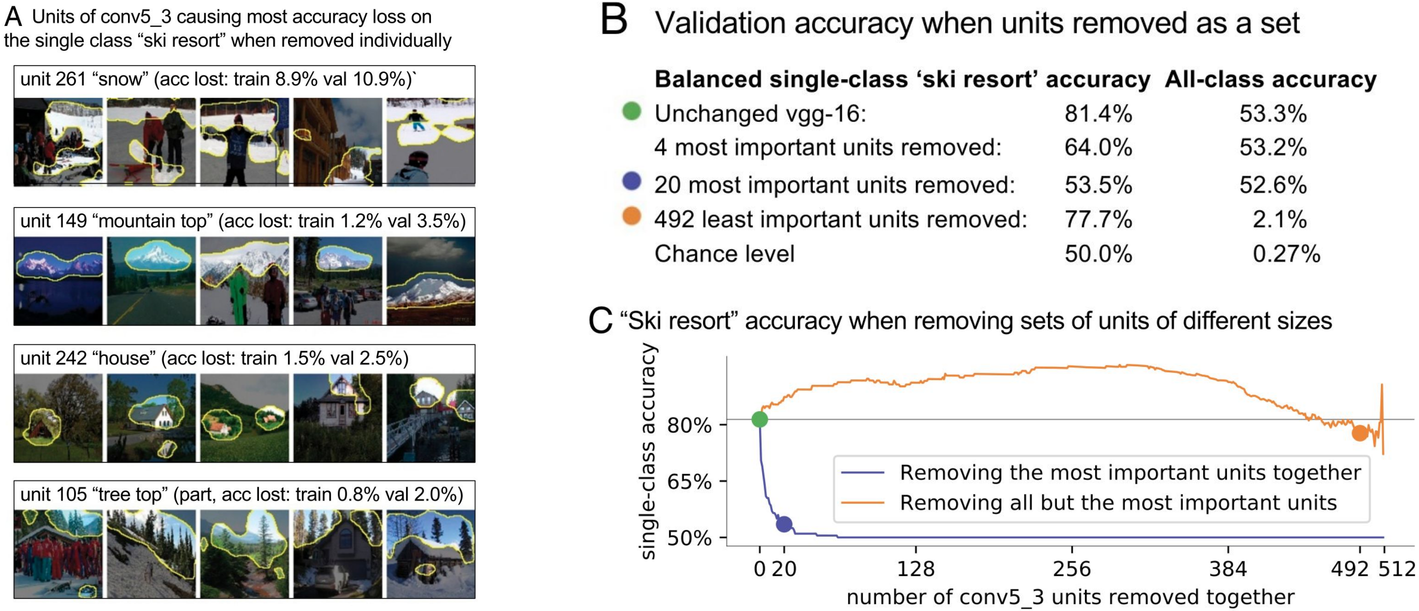
- Object detector units **automatically** emerge when train with larger scene classification
- Training with larger scene classification will results in units that match with **different levels** of concepts (objects, parts, materials, colors)
- Those units are not concepts exclusive
- Units that have IoU < 4% is excluded



Highlighted regions are those whose activation is among 1% quantile of total activation of that unit

Results - how important is each units causally?

- Removing important units (to zero, ranked by IoU) **hurts** the classification badly
- Measure accuracy by balanced binary classification for individual classes
- Removing redundant units **even help improve** the accuracy



A Units of conv5_3 causing most accuracy loss on the single class “ski resort” when removed individually

unit 261 “snow” (acc lost: train 8.9% val 10.9%)`

unit 149 “mountain top” (acc lost: train 1.2% val 3.5%)

unit 242 “house” (acc lost: train 1.5% val 2.5%)

unit 105 “tree top” (part, acc lost: train 0.8% val 2.0%)

B Validation accuracy when units removed as a set

	Balanced single-class ‘ski resort’ accuracy	All-class accuracy
● Unchanged vgg-16:	81.4%	53.3%
4 most important units removed:	64.0%	53.2%
● 20 most important units removed:	53.5%	52.6%
● 492 least important units removed:	77.7%	2.1%
Chance level	50.0%	0.27%

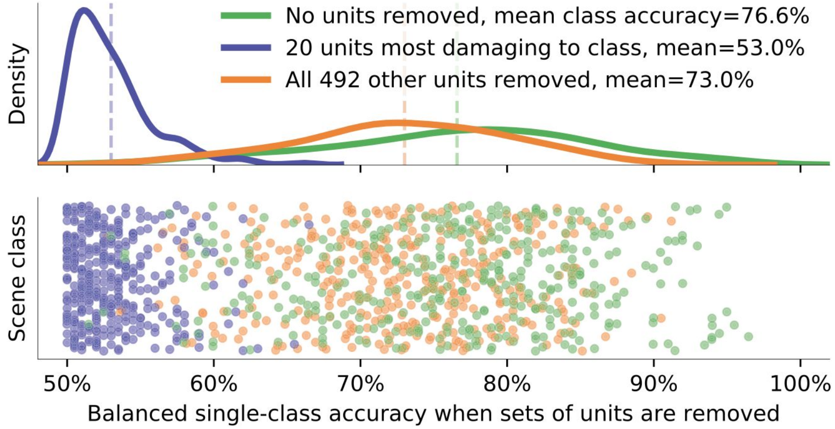
C “Ski resort” accuracy when removing sets of units of different sizes

single-class accuracy
80%
65%
50%
0 20
128
256
384
492 512
number of conv5_3 units removed together
Removing the most important units together
Removing all but the most important units

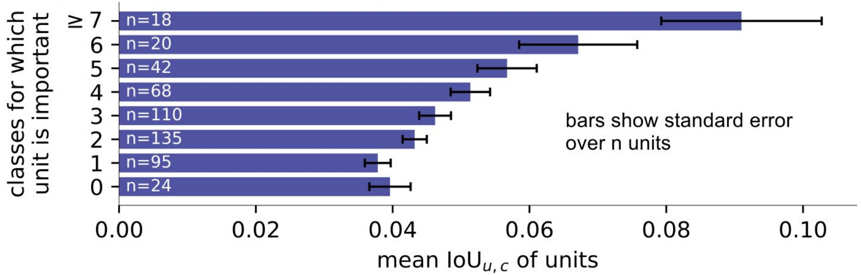
Results - how important is each units causally?

- Removing 20 most important units hurt most whereas others barely impact acc
- Units shared by multiple classes has higher IoU (more interpretable)

D Removing 20 most- and 492 least-important units for all scene classes

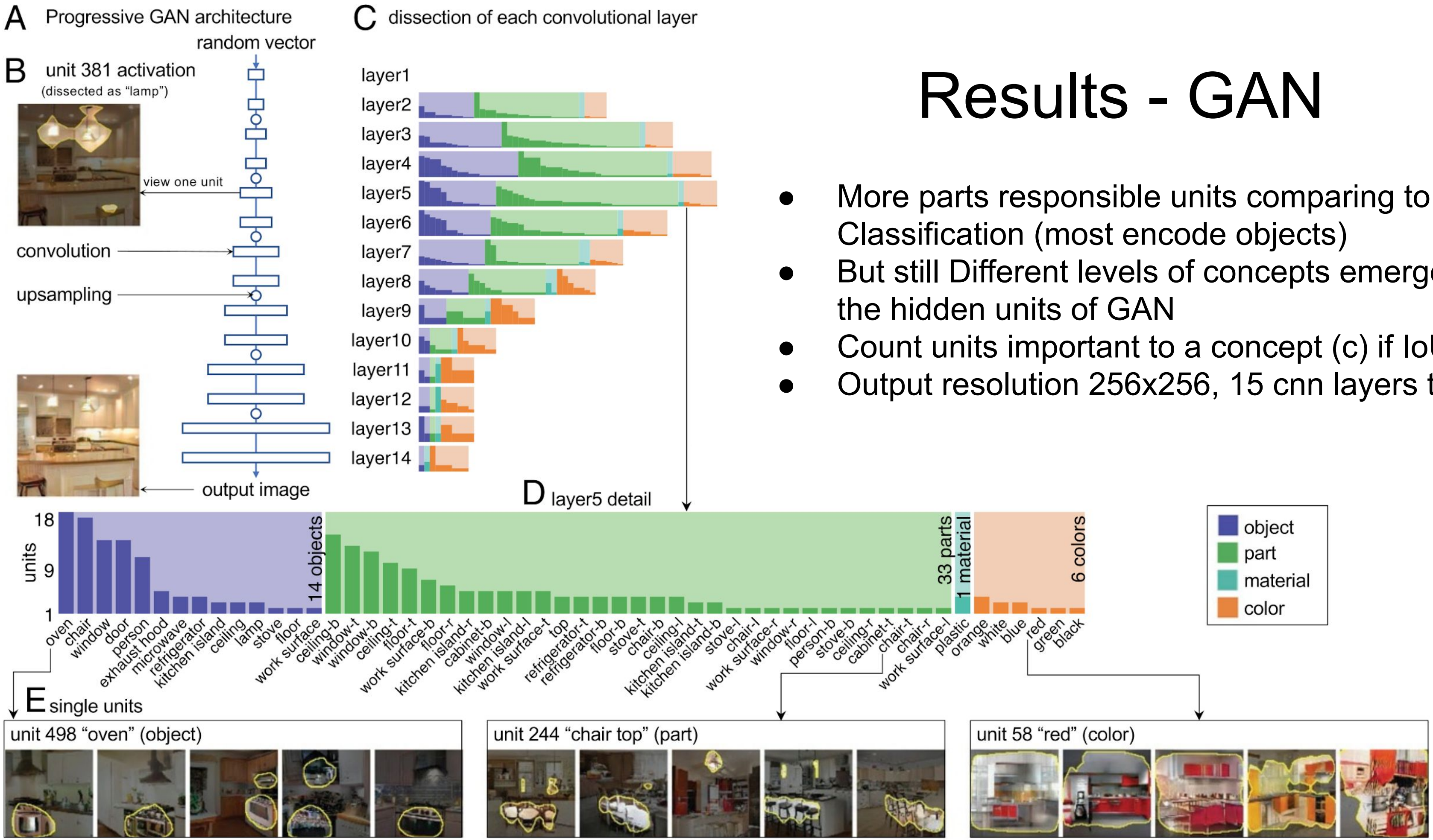


E Interpretability of units that are important to more and fewer classes



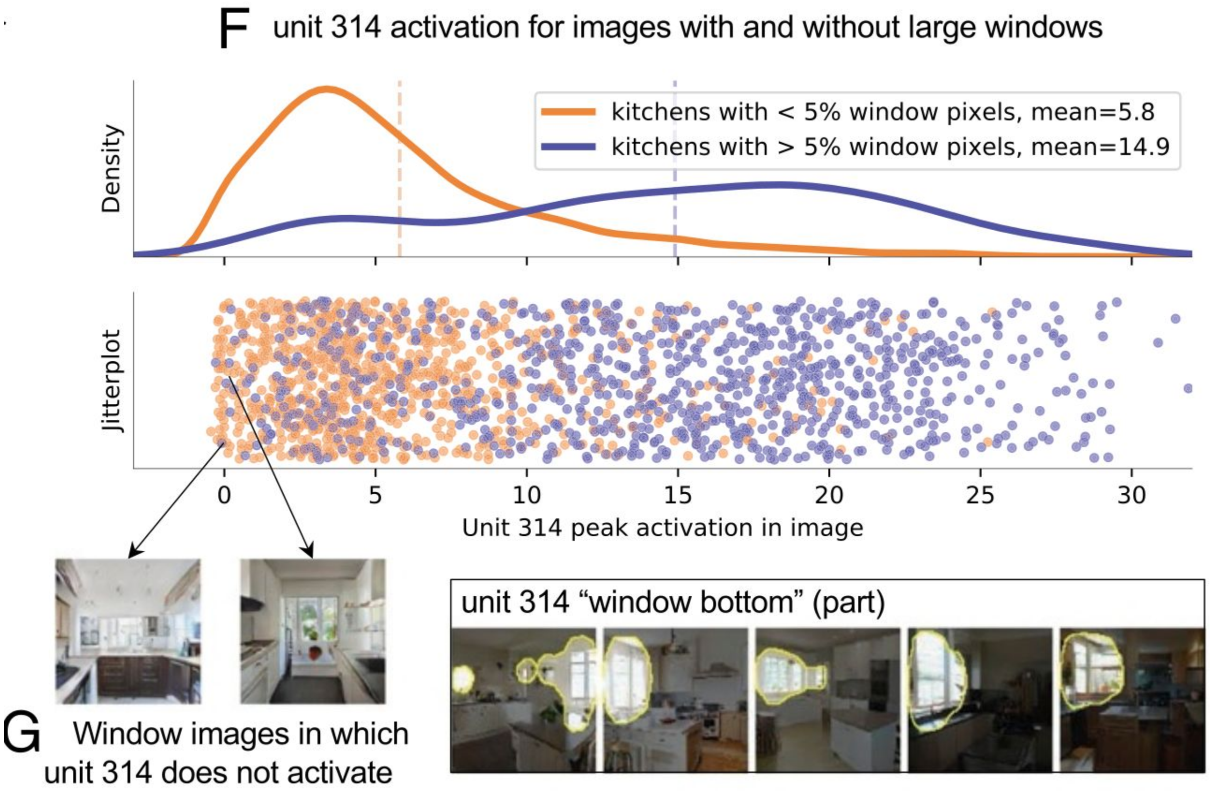
Results - GAN

- More parts responsible units comparing to Classification (most encode objects)
- But still Different levels of concepts emerges in the hidden units of GAN
- Count units important to a concept (c) if IoU > 4%
- Output resolution 256x256, 15 cnn layers total



Results - GAN

- Concepts (c) related units (u) in GAN
- Using window specific unit activation can classify if generated image has window (78.2% Accuracy)
- But there are counter-examples in Figure G

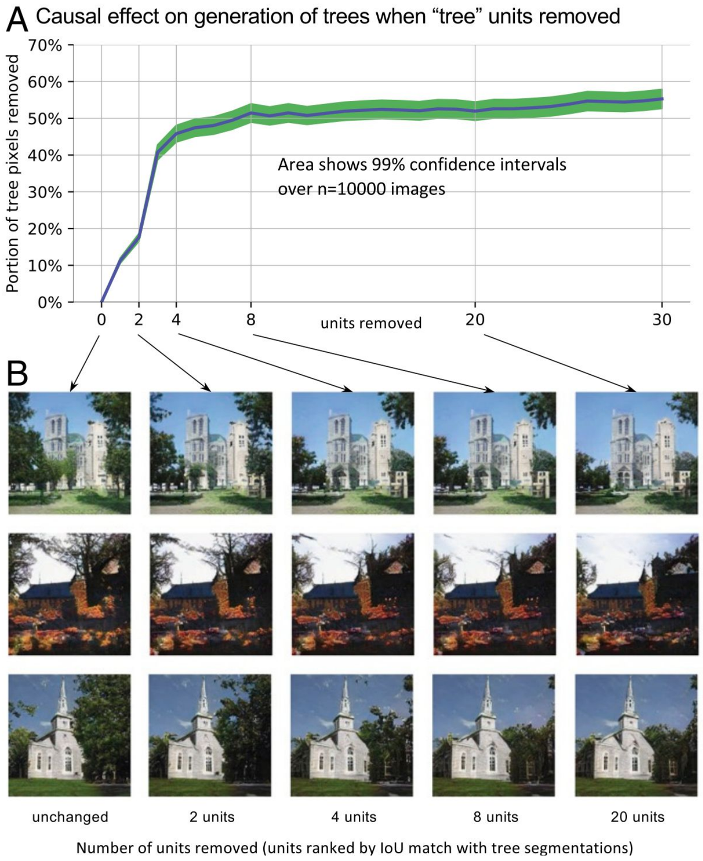


F unit 314 activation for images with and without large windows

G Window images in which unit 314 does not activate

Results - Causal role of units in GAN (remove)

- Removing tree specific units (layer 4) results in tree removal in the generated image
- Tree pixels are identified by segmentation network
- Remove tree units leave the whole image intact
- ! Remove the tree units even reveal the church which were occluded before -> suggesting the network compute compositional structures.

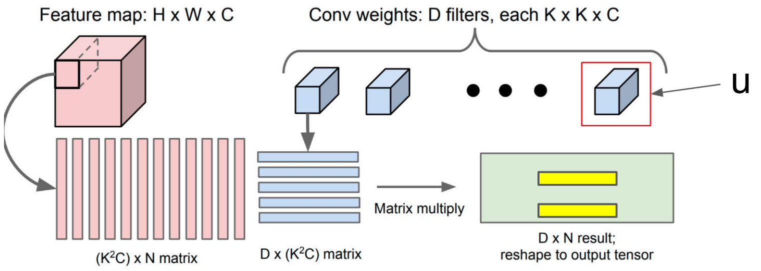


A Causal effect on generation of trees when “tree” units removed

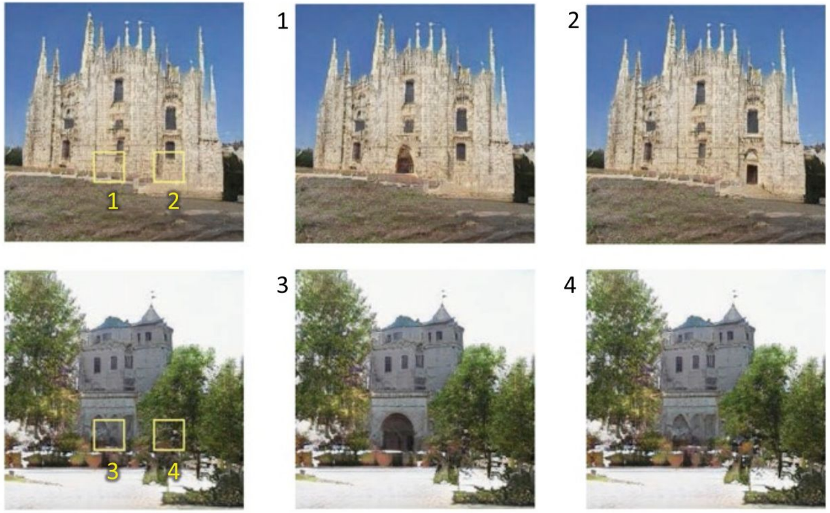
B

Results - Causal role of units in GAN (activation)

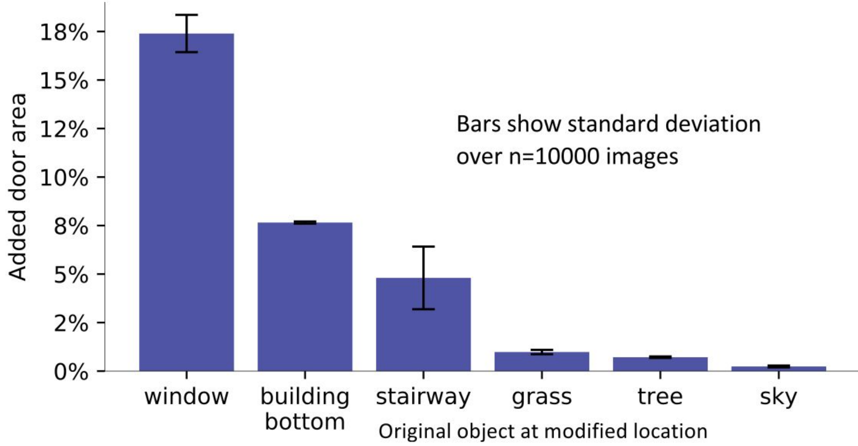
- Activating “door” units at certain location
- Adding depends on the context of the location



C Effect of activating “door” units depends on location



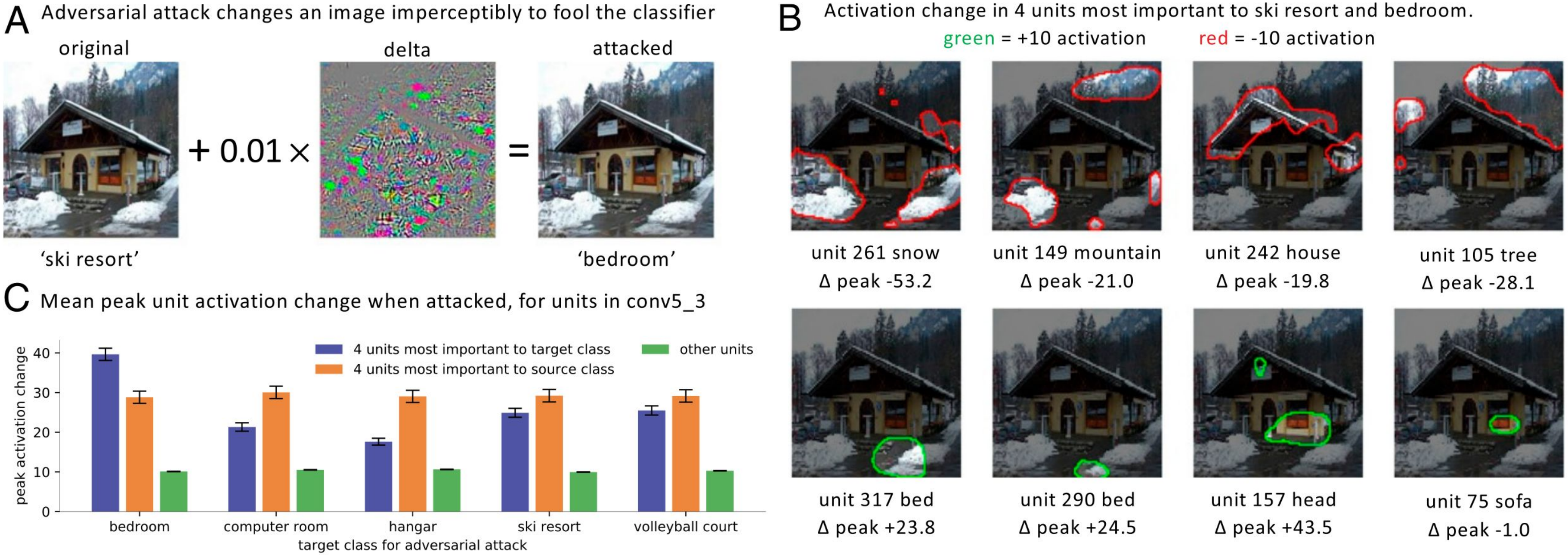
D Effect of activating “door” units depends on object context



Bars show standard deviation over n=10000 images

Application: Analyzing Adversarial Attack

- Adversarial attack diminishes the firing of important units for original class
- However, increase the firing of the important units for the target class



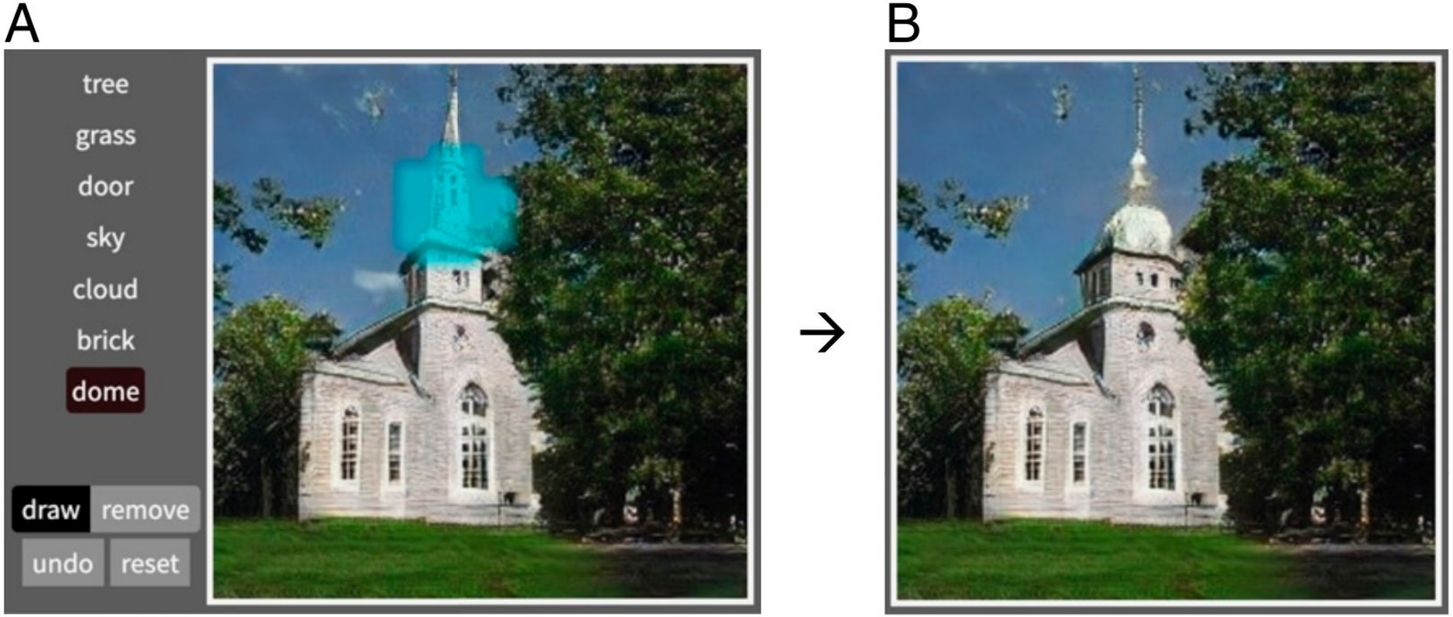
A Adversarial attack changes an image imperceptibly to fool the classifier

B Activation change in 4 units most important to ski resort and bedroom.
green = +10 activation red = -10 activation

C Mean peak unit activation change when attacked, for units in conv5_3

Application: GAN concept painting

- Activate the important 20 units of selected concept at certain location results in painting the concept.


A
tree
grass
door
sky
cloud
brick
dome
draw
remove
undo
reset
→
B