CS839 Special Topics in Deep Learning

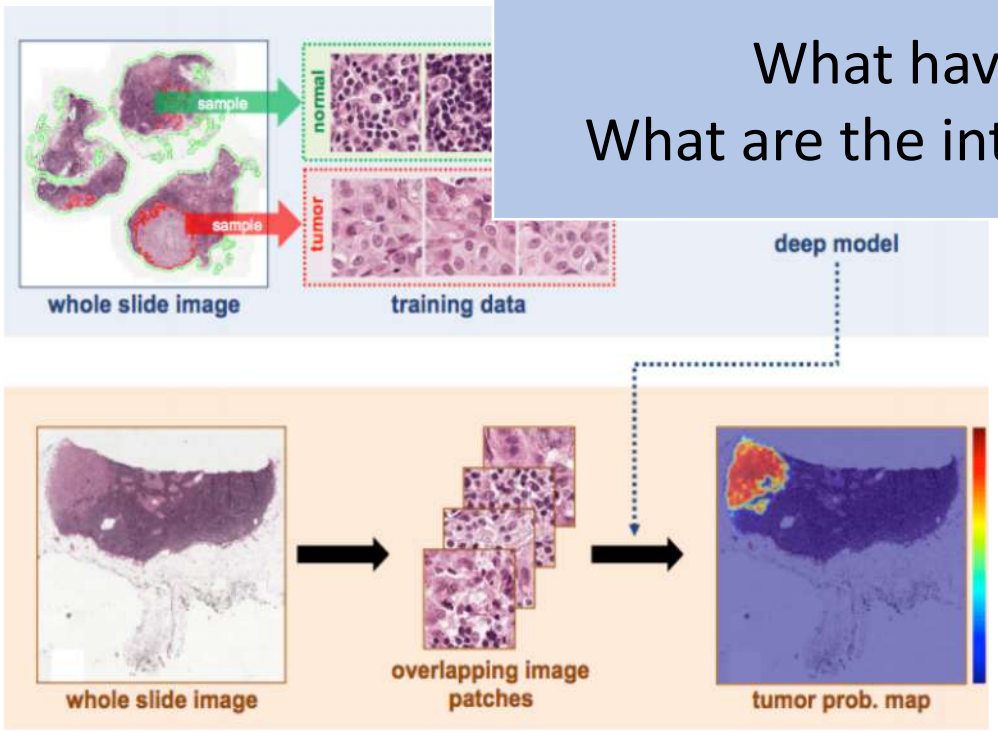
Introduction on Interpretable Deep Learning

Yiyou Sun
University of Wisconsin-Madison

October 8, 2020

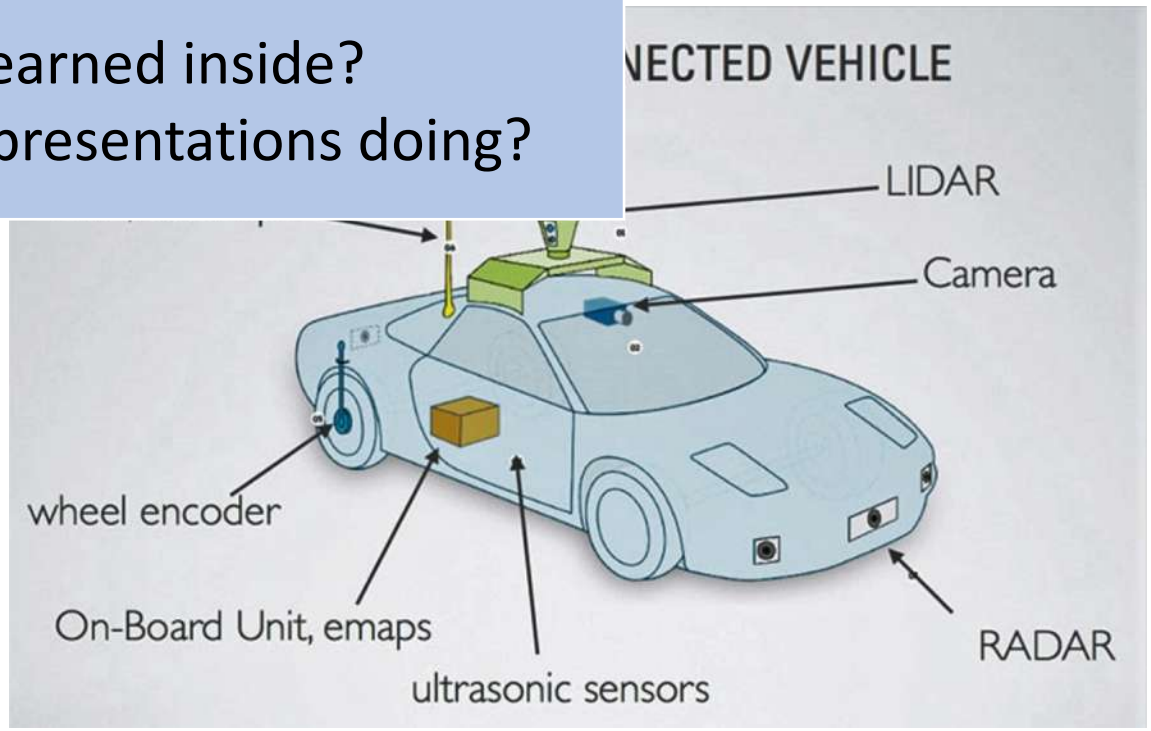
Deep Neural Networks are Everywhere

Making Medical Decision



What have been learned inside?
What are the internal representations doing?

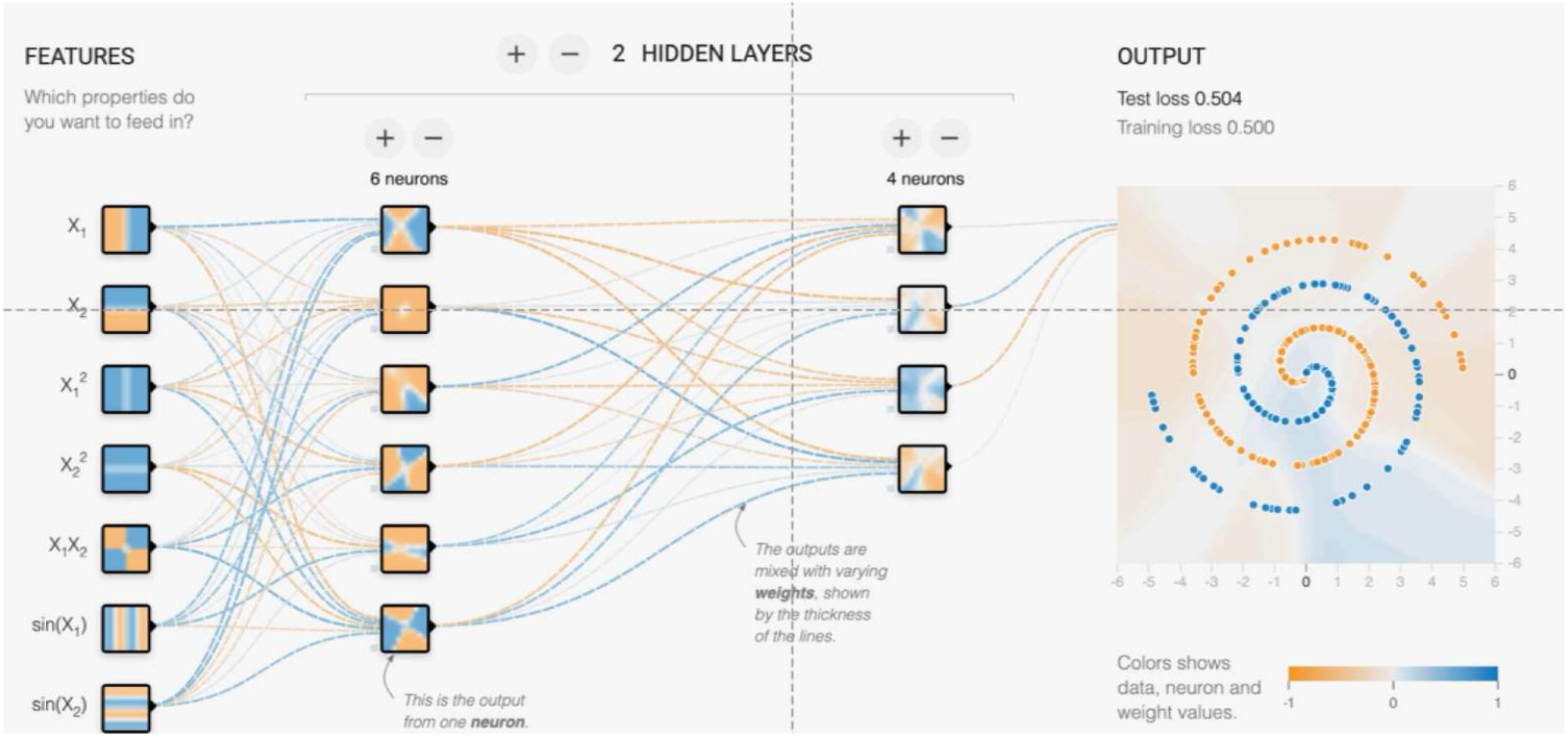
Self-Driving Car



Outline

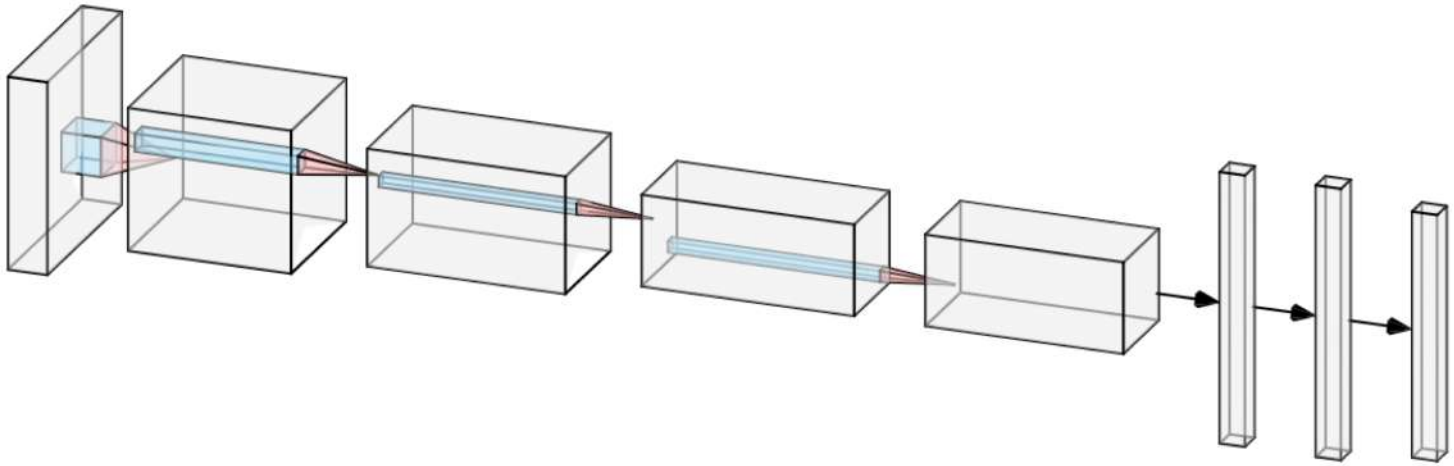
1. **Visualization Methods**
2. Attribution Methods
3. Interpretable System
4. Interpretation for GAN

Let’s Start with a toy demo!



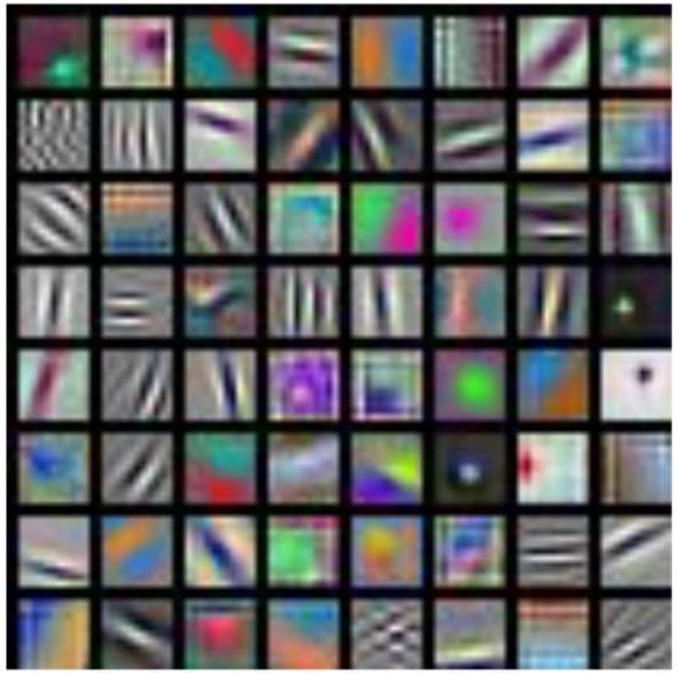
[1] https://playground.tensorflow.org/

What about a deep CNN model?



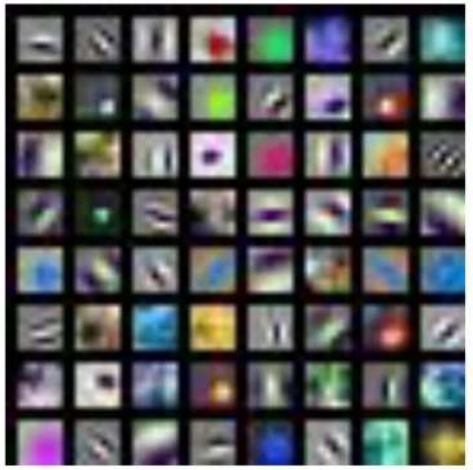
[1] http://alexlenail.me/NN-SVG/AlexNet.html

We can print first layer's filters!



AlexNet:
64 x 3 x 11 x 11

ResNet-18:
64 x 3 x 7 x 7



ResNet-101:
64 x 3 x 7 x 7

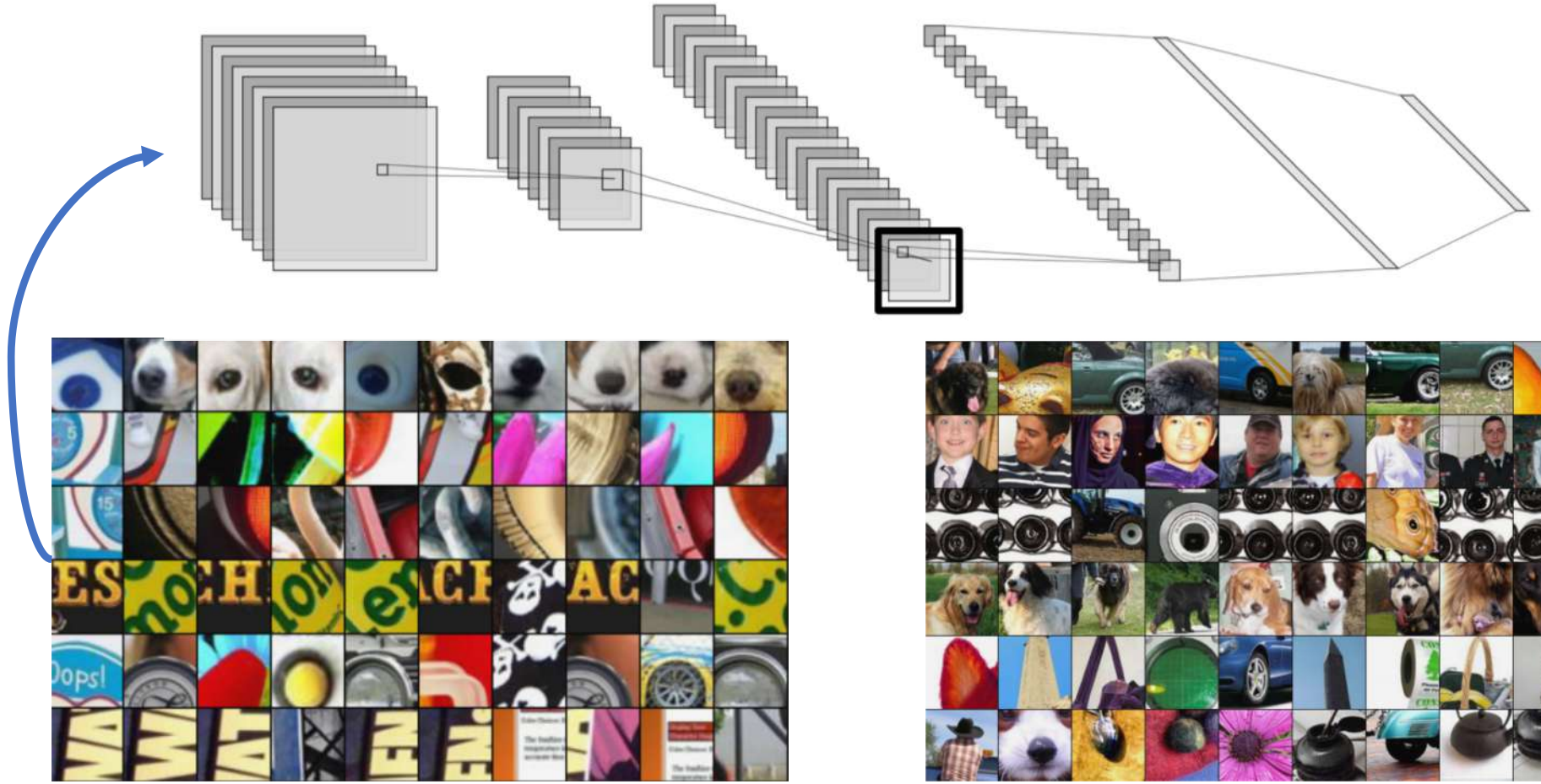
DenseNet-121:
64 x 3 x 7 x 7

[1] "Visualizing and Understanding", cs231n-2017 lecture 12, Fei-Fei Li, Justin Johnson, Serena Yeung

What about filters in deep layers?

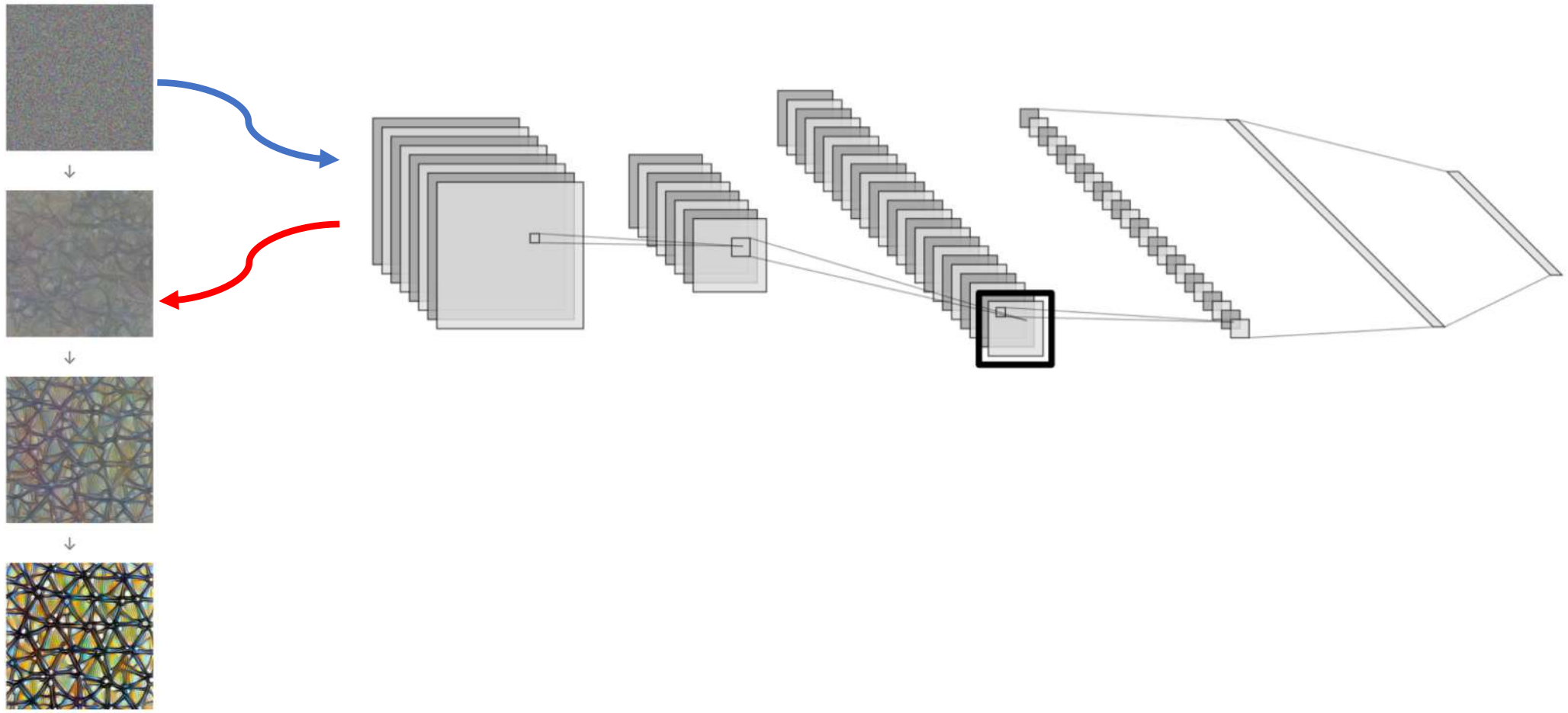
[1] https://cs231n.github.io/understanding-cnn/

Maximally Activated Patches



[1] https://arxiv.org/pdf/1412.6806.pdf

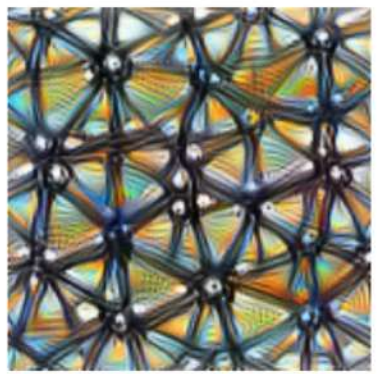
Deep Feature Visualization by Gradient Ascent



[1] https://distill.pub/2017/feature-visualization/

Neuron

`layer`$_n$`[x,y,z]`

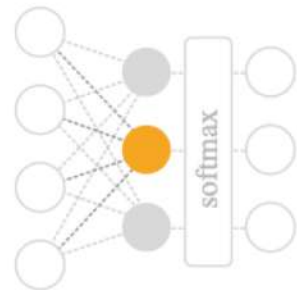


Channel

`layer`$_n$`[:,:,z]`

Layer/DeepDream

`layer`$_n$`[:,:,:]`^{2}



Class Logits

`pre_softmax[k]`

Class Probability

`softmax[k]`

Tricks to Make Natural Visualization

- **Frequency penalization**

- Penalize variance

- Penalize high-freq

- **Transformatio**

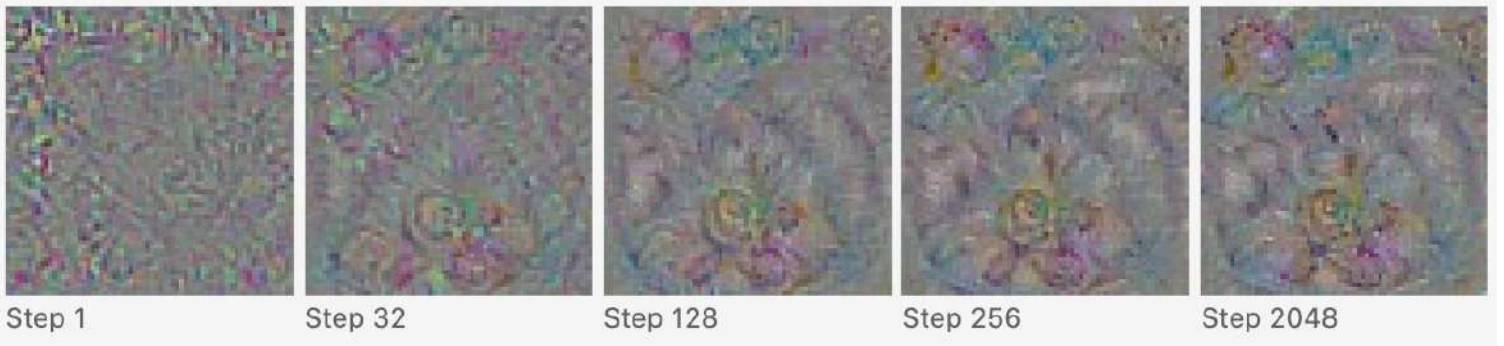
- Generate images that still activate the optimization target even with **jitter**, **rotation** or **scaling**

- **Learned priors**

- learn a model(GAN, VAE) of the real data to generate photorealistic visualizations

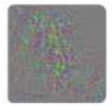
Unregularized

Frequency Penalization



Transformation robustness

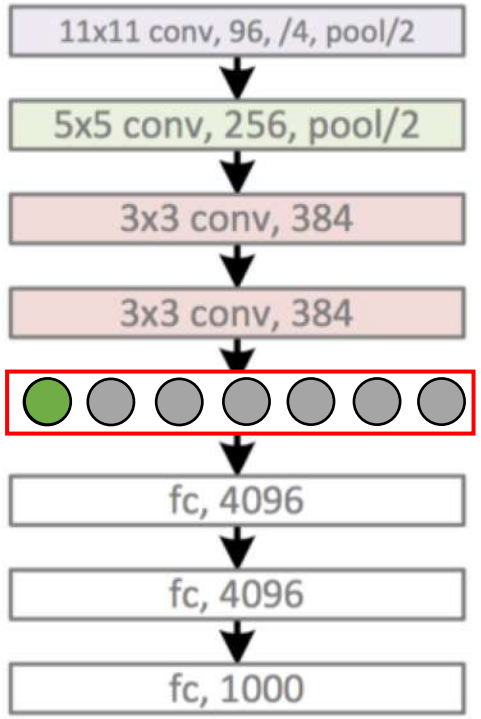
[1] https://distill.pub/2017/feature-visualization/

	Unregularized	Frequency Penalization	Transformation Robustness	Learned Prior	Dataset Examples
 **Erhan, *et al.*, 2009** [3] Introduced core idea. Minimal regularization.	■				
 **Szegedy, *et al.*, 2013** [11] Adversarial examples. Visualizes with dataset examples.	■				■
 **Mahendran & Vedaldi, 2015** [7] Introduces total variation regularizer. Reconstructs input from representation.		■			
 **Nguyen, *et al.*, 2015** [14] Explores counterexamples. Introduces image blurring.		■			
 **Mordvintsev, *et al.*, 2015** [4] Introduced jitter & multi-scale. Explored GMM priors for classes.			■	■	
 **Øygard, *et al.*, 2015** [15] Introduces gradient blurring. (Also uses jitter.)		■	■		
 **Tyka, *et al.*, 2016** [16] Regularizes with bilateral filters. (Also uses jitter.)		■	■		
 **Mordvintsev, *et al.*, 2016** [17] Normalizes gradient frequencies. (Also uses jitter.)		■	■		
 **Nguyen, *et al.*, 2016** [18] Paramaterizes images with GAN generator.				■	
 **Nguyen, *et al.*, 2016** [10] Uses denoising autoencoder prior to make a generative model.				■	

[1] https://distill.pub/2017/feature-visualization/

Comparison

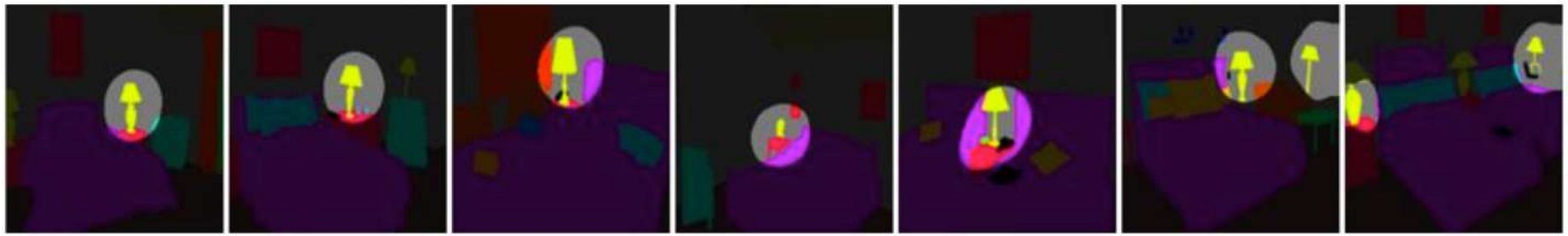
From Visualization to Interpretation



Top Activated Images

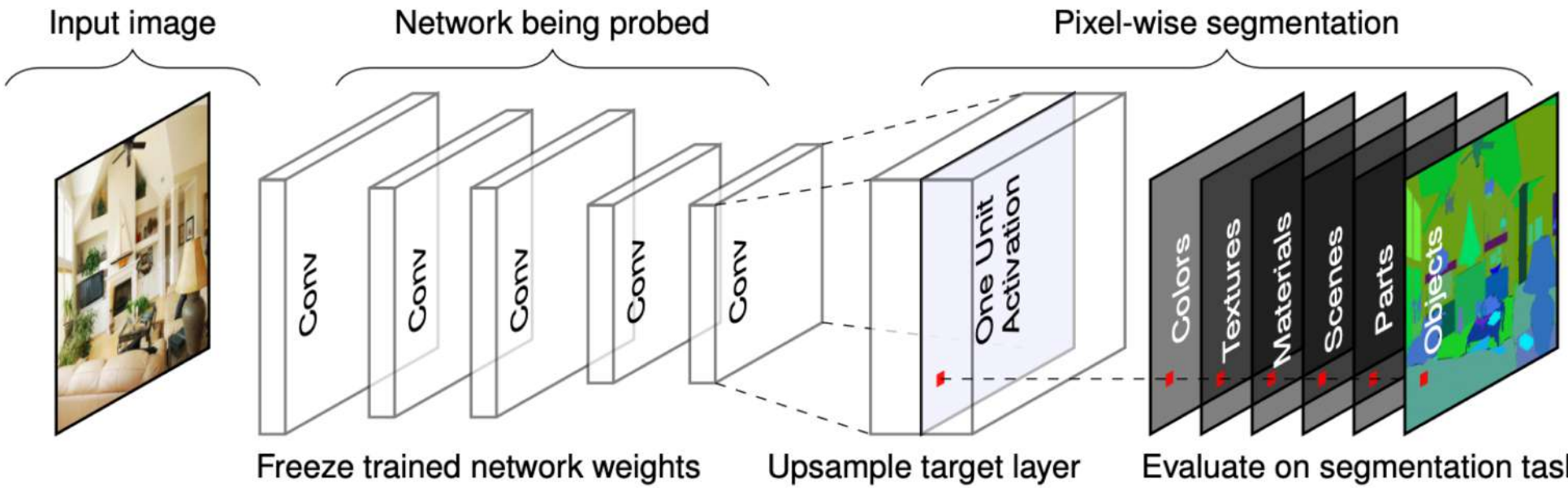
Unit 1

Lamp Intersection over Union (IoU)= 0.12



[1] Zhou et al., Network Dissection. PAMI 2018.

IoU Score Calculation



$$IoU(Unit, Concept) = \frac{\sum |Area(Unit) \cap Area(Concept)|}{\sum |Area(Unit) \cup Area(Concept)|}$$

[1] Zhou et al., Network Dissection. PAMI 2018.

Broadly and Densely (**Broden**) Annotated 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

Describable Textures
Cimpoi et al, CVPR'14

Colors

Total = 63,305 images
1,197 visual concepts

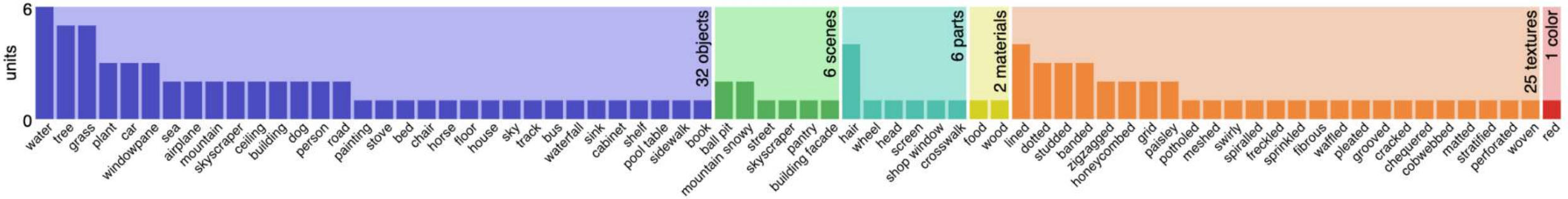
[1] Zhou et al., Network Dissection. PAMI 2018.

AlexNet-Places205 conv5 unit 138: heads

AlexNet-Places205 conv5 unit 215: castles

AlexNet-Places205 conv5 unit 13: lamps

AlexNet-Places205 conv5 unit 53: stairways

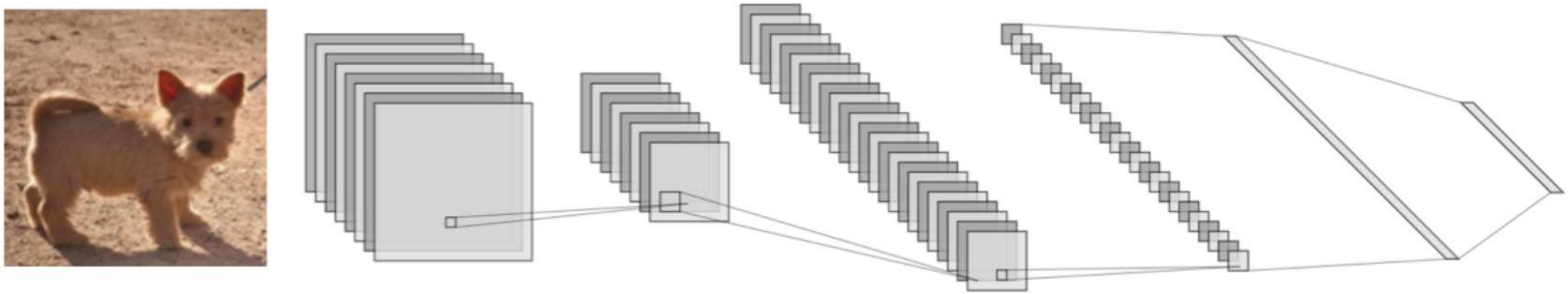


[1] Zhou et al., Network Dissection. PAMI 2018.

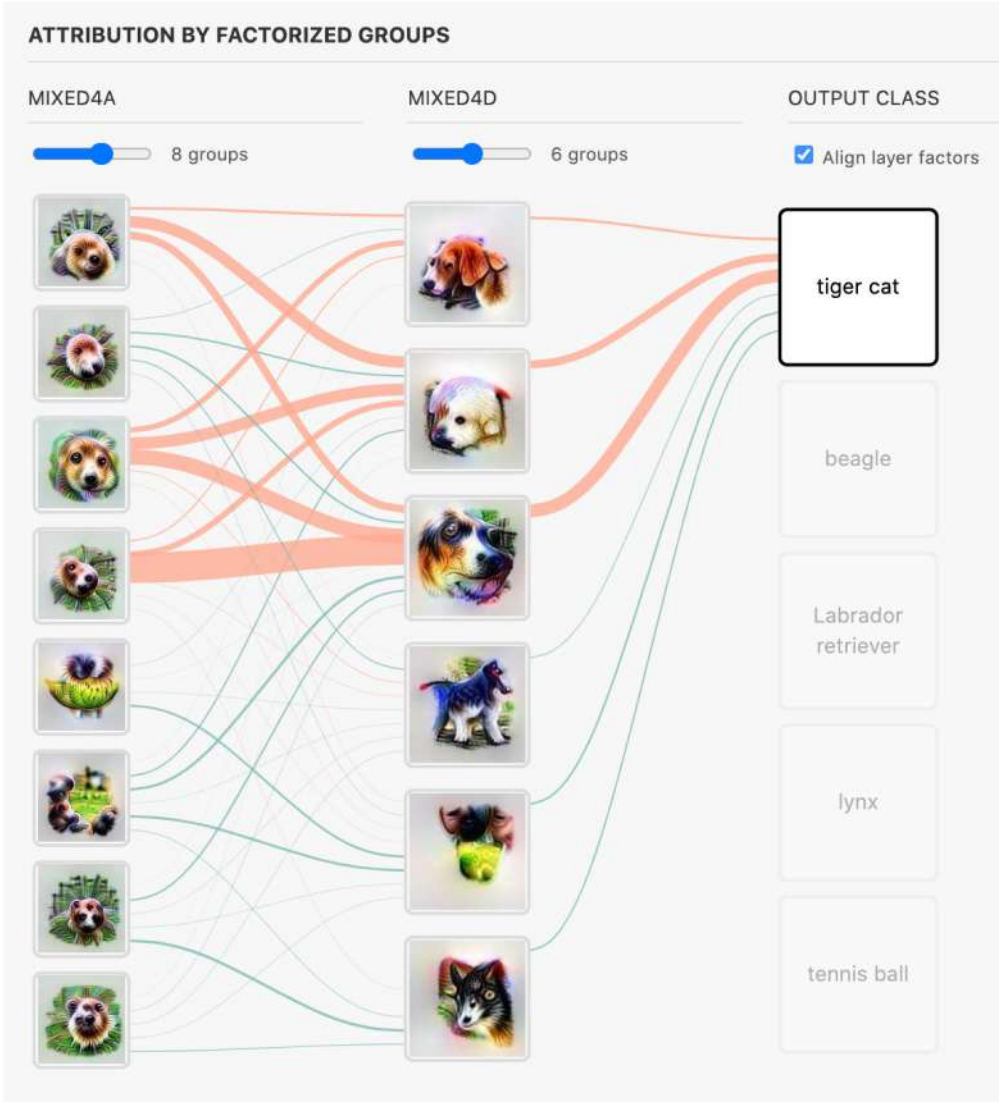
Outline

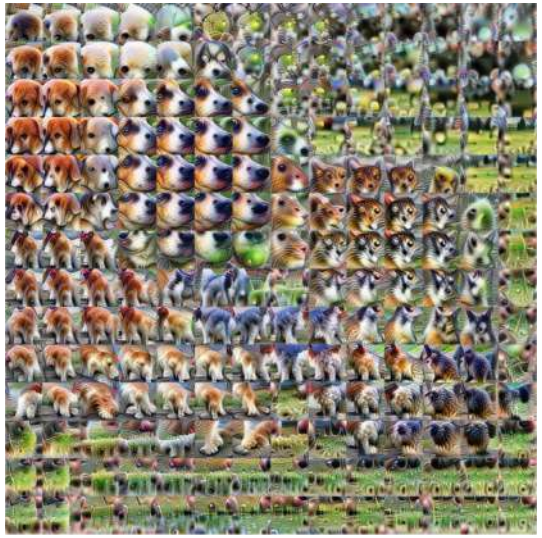
1. Visualization Methods
2. **Attribution Methods**
3. Interpretable System
4. Interpretation for GAN

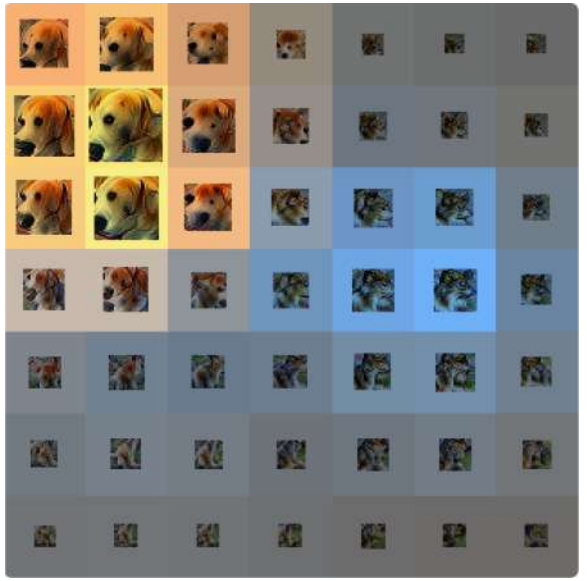
Which pixels matter for classification?



The Building Blocks of Interpretability

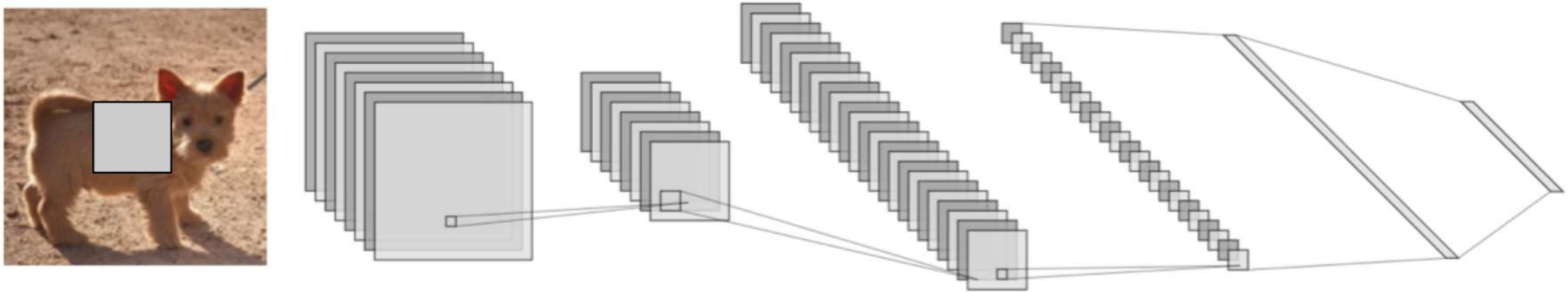




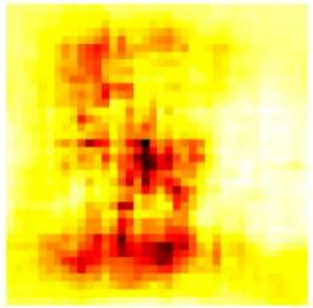


[1] https://distill.pub/2018/building-blocks/

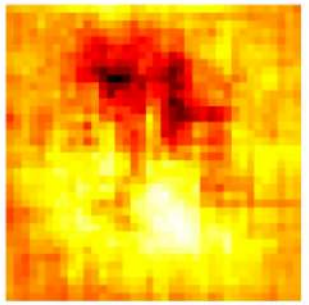
Occlusion Based Methods



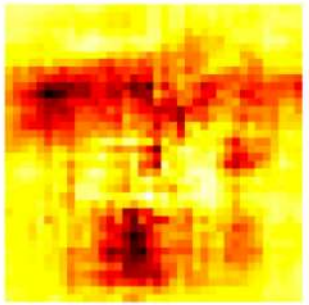
schooner



African elephant, Loxodonta africana

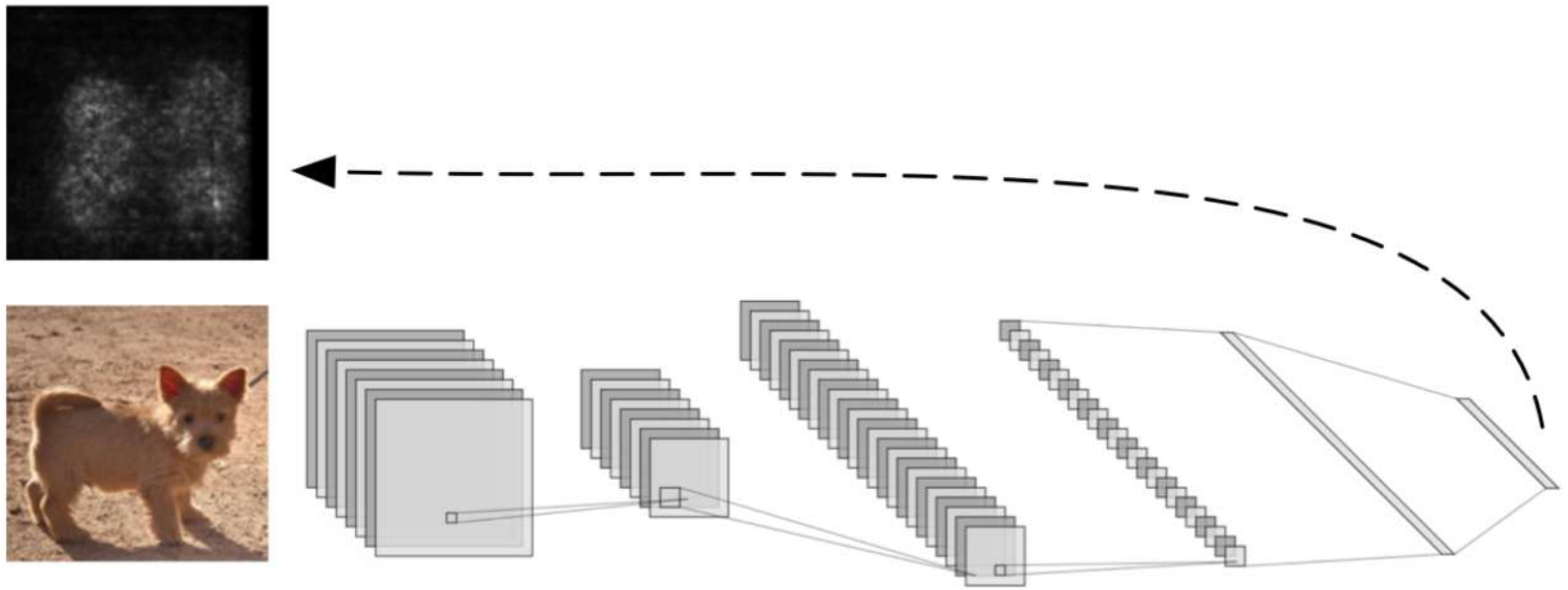


go-kart



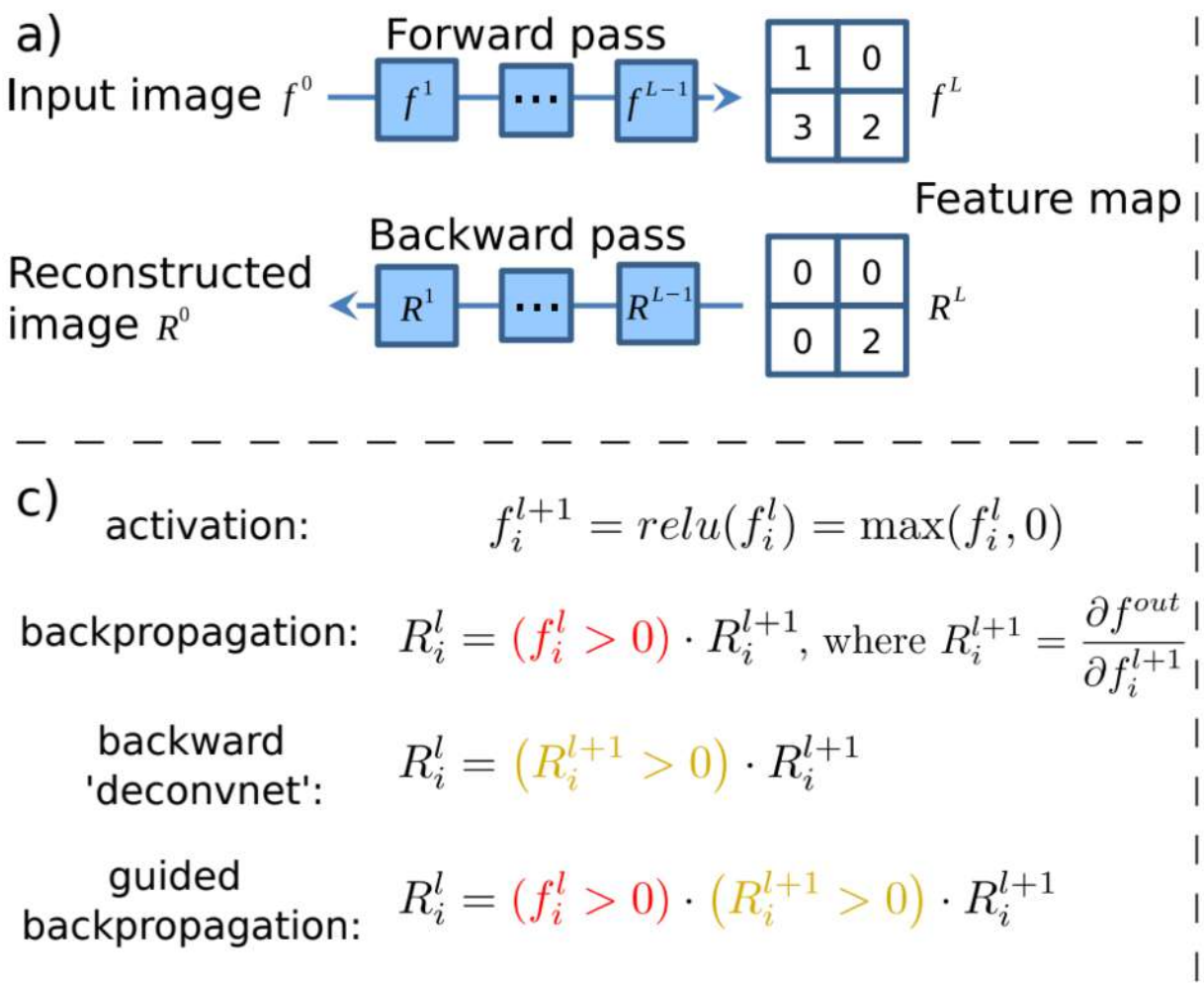
[1] Zeiler and Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV 2014

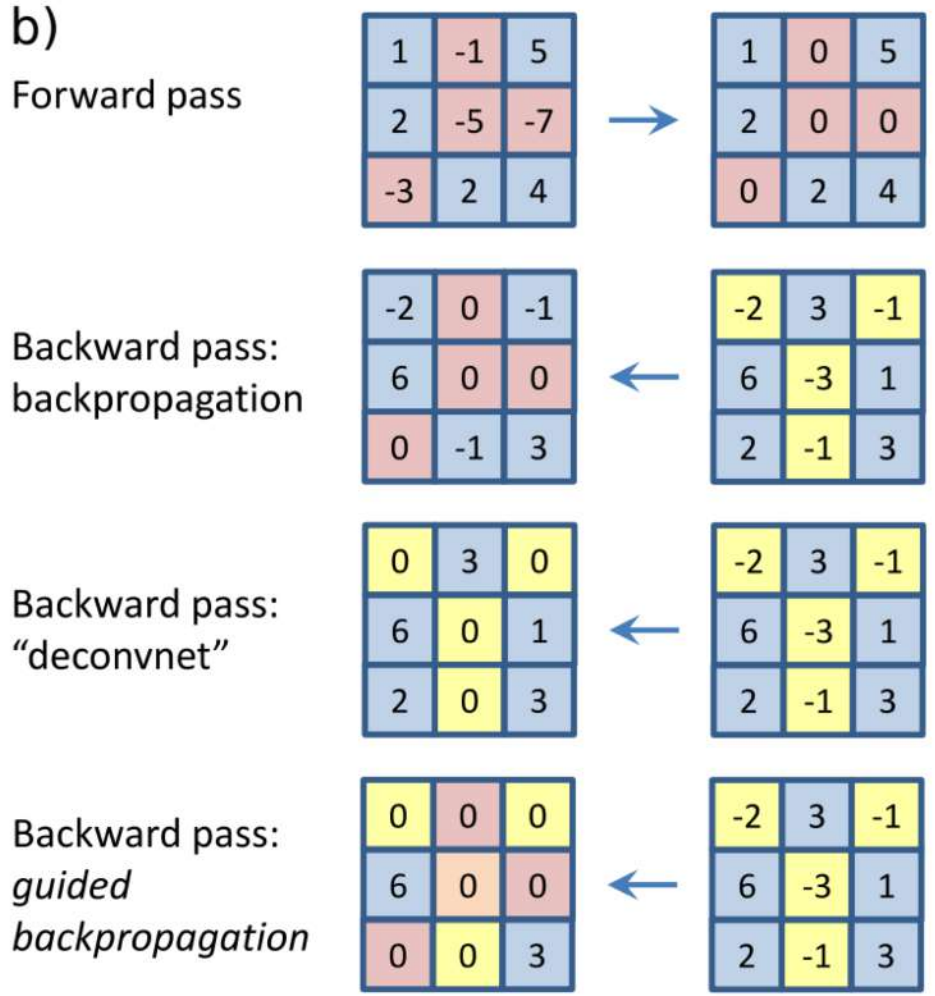
Saliency Map: Gradient Visualization



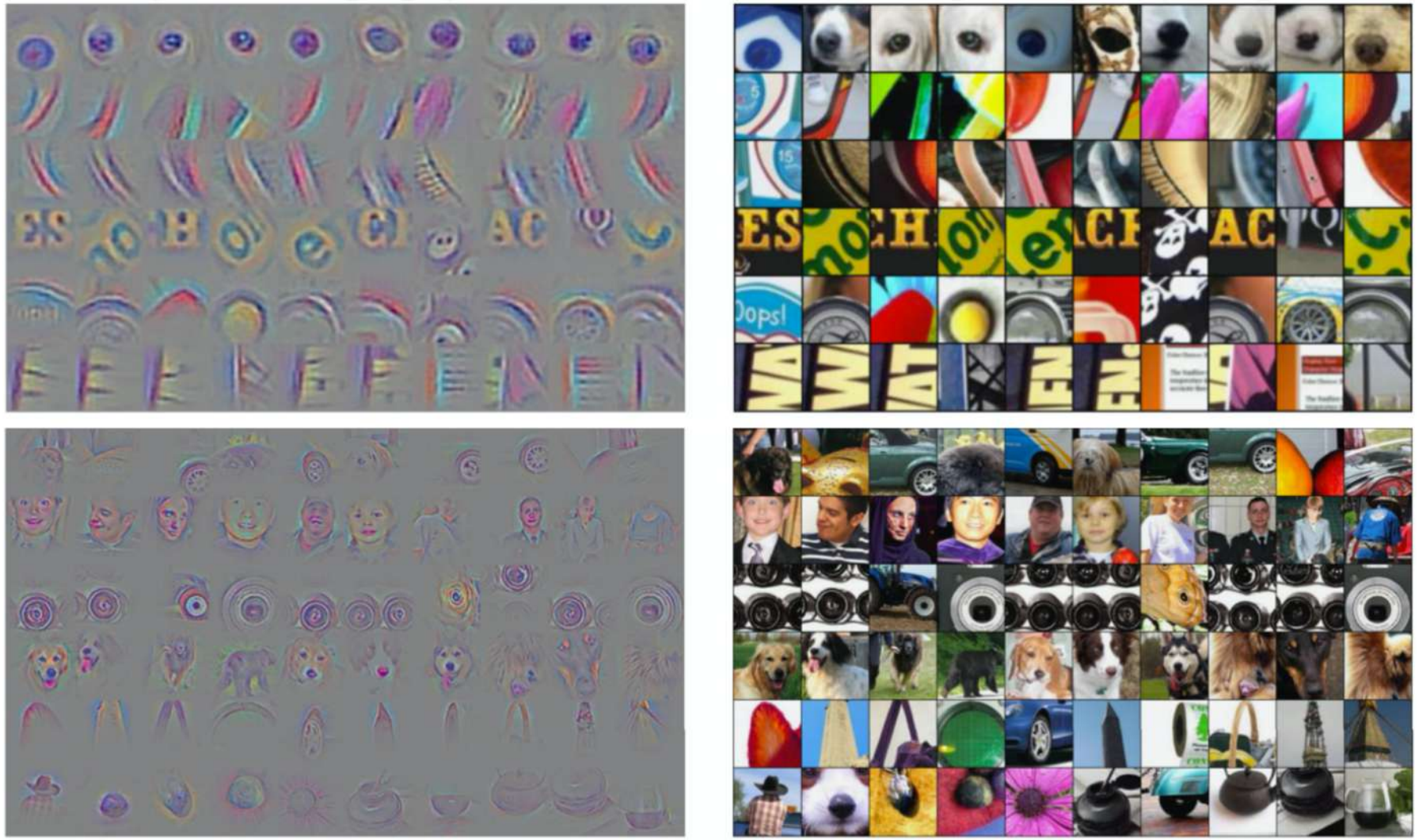
[1] Simonyan, Vedaldi, and Zisserman, “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”, ICLR Workshop 2014.

Guided Backprop





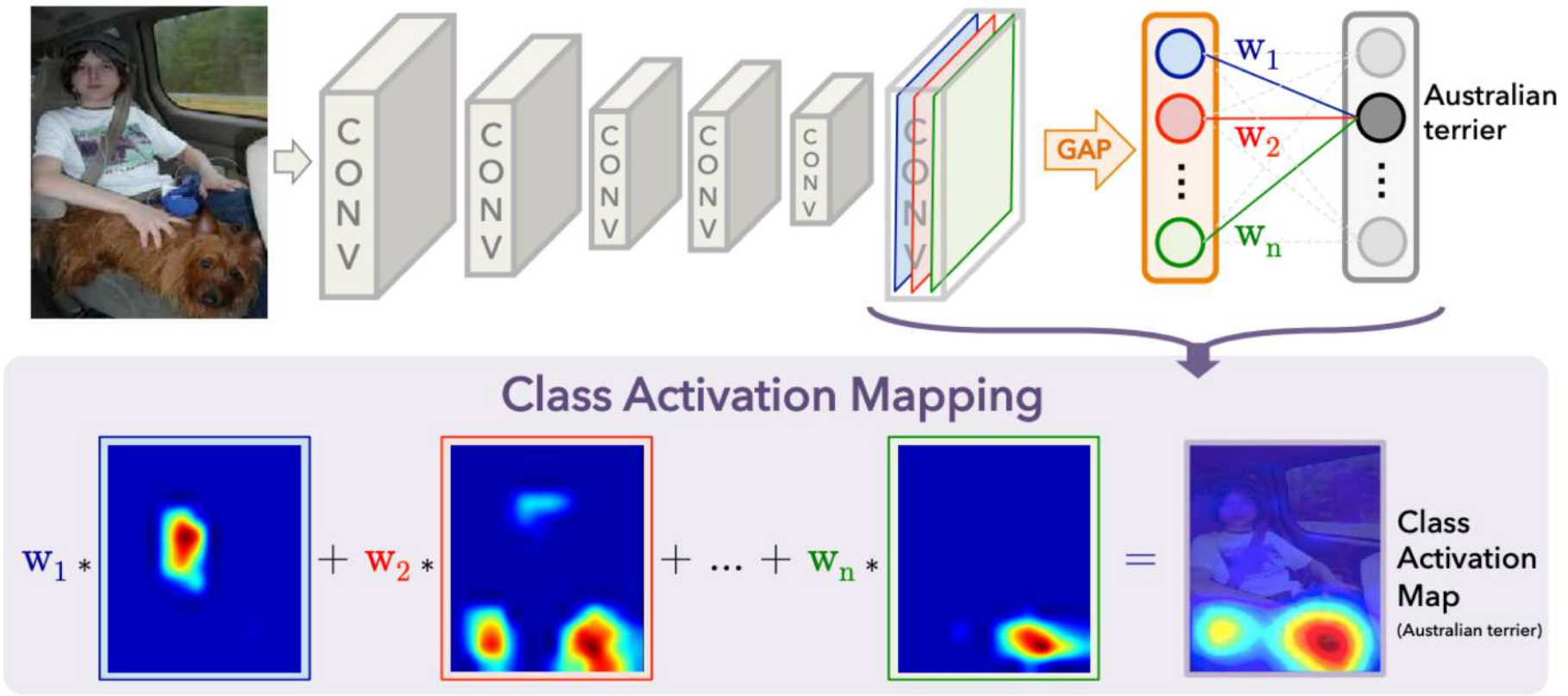
[1] Springenberg et al, "Striving for Simplicity: The All Convolutional Net", ICLR Workshop 2015



[1] Springenberg et al, “Striving for Simplicity: The All Convolutional Net”, ICLR Workshop 2015

CAM

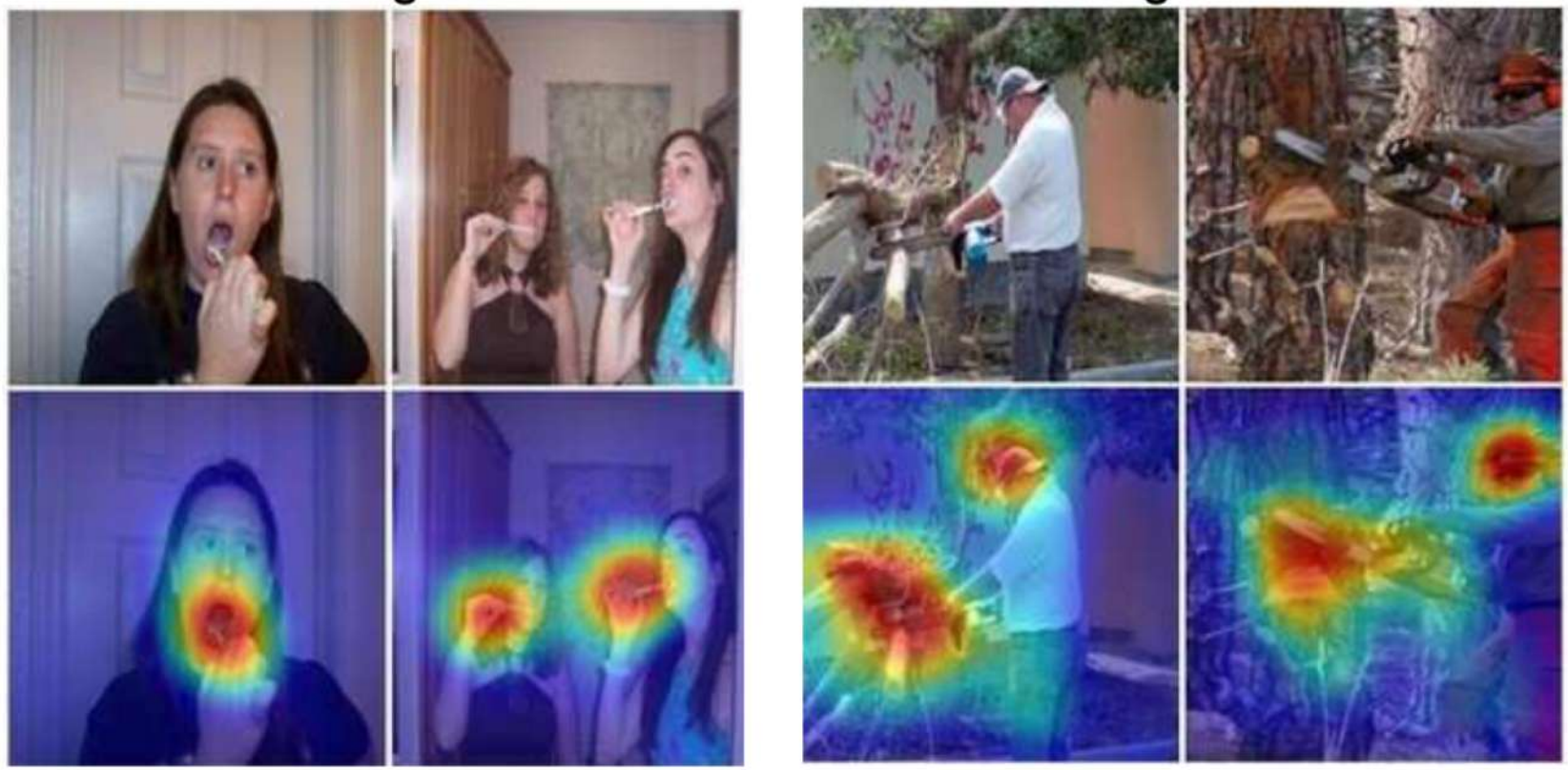
Can we use last feature map as attribution?



[1] Bolei et al, “Learning Deep Features for Discriminative Localization”, CVPR 2016

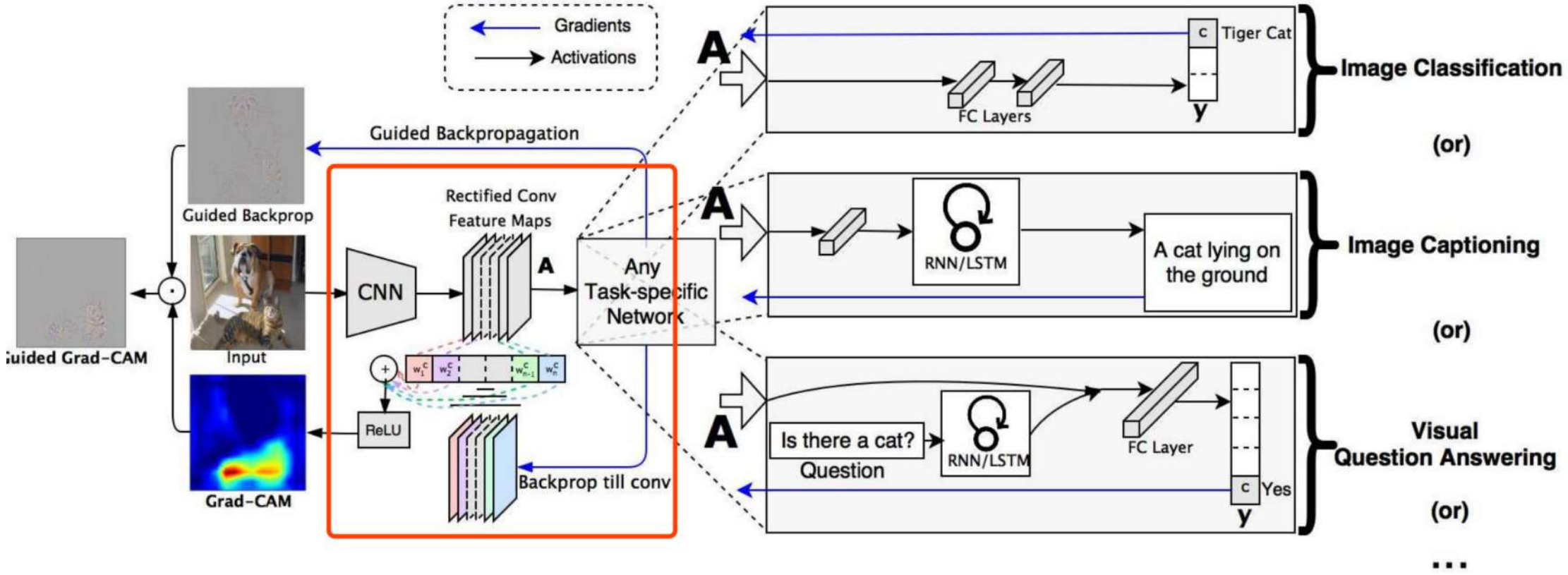
Brushing teeth

Cutting trees



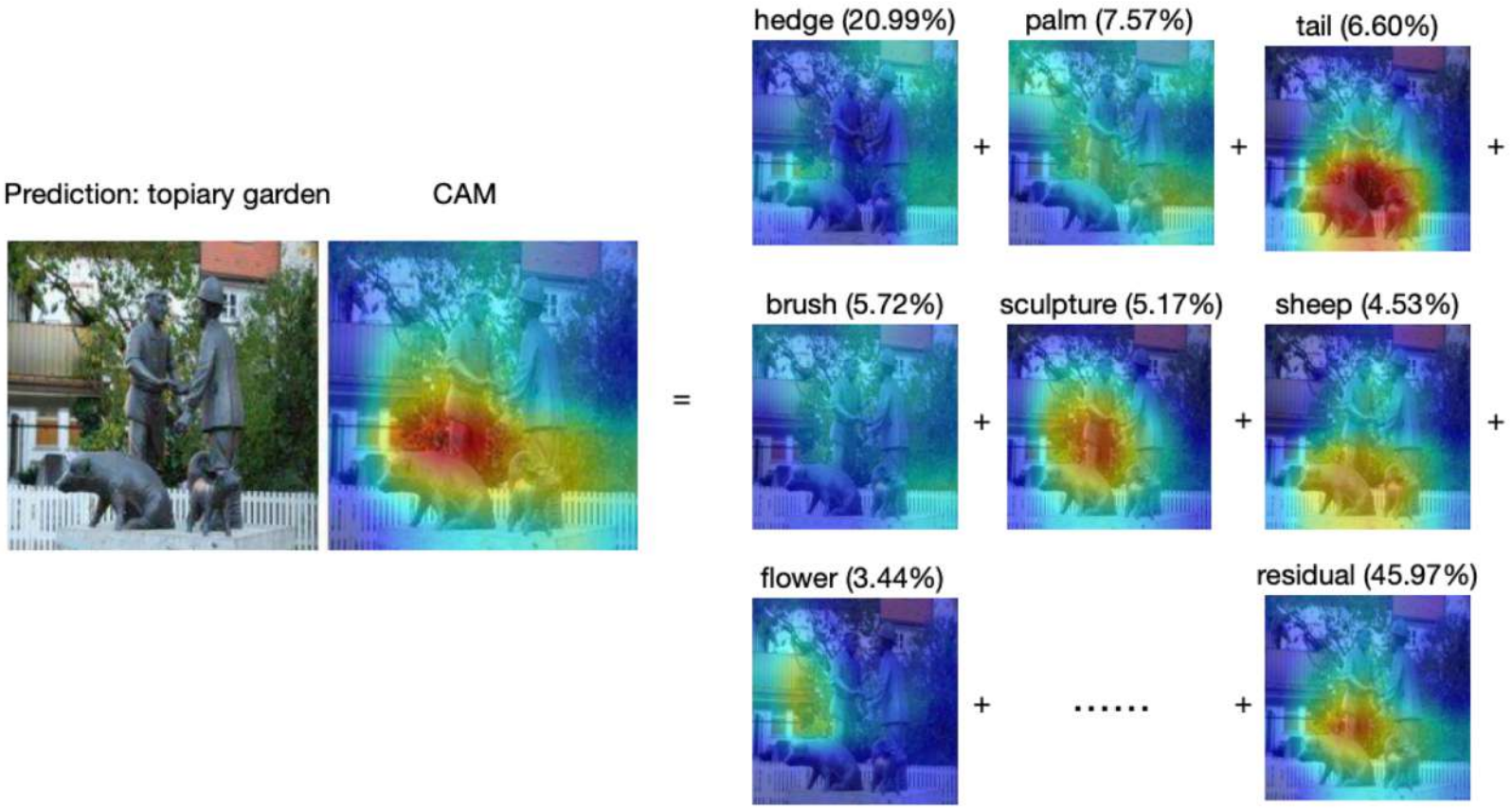
Grad-CAM

Can we use **intermediate** feature map as attribution?

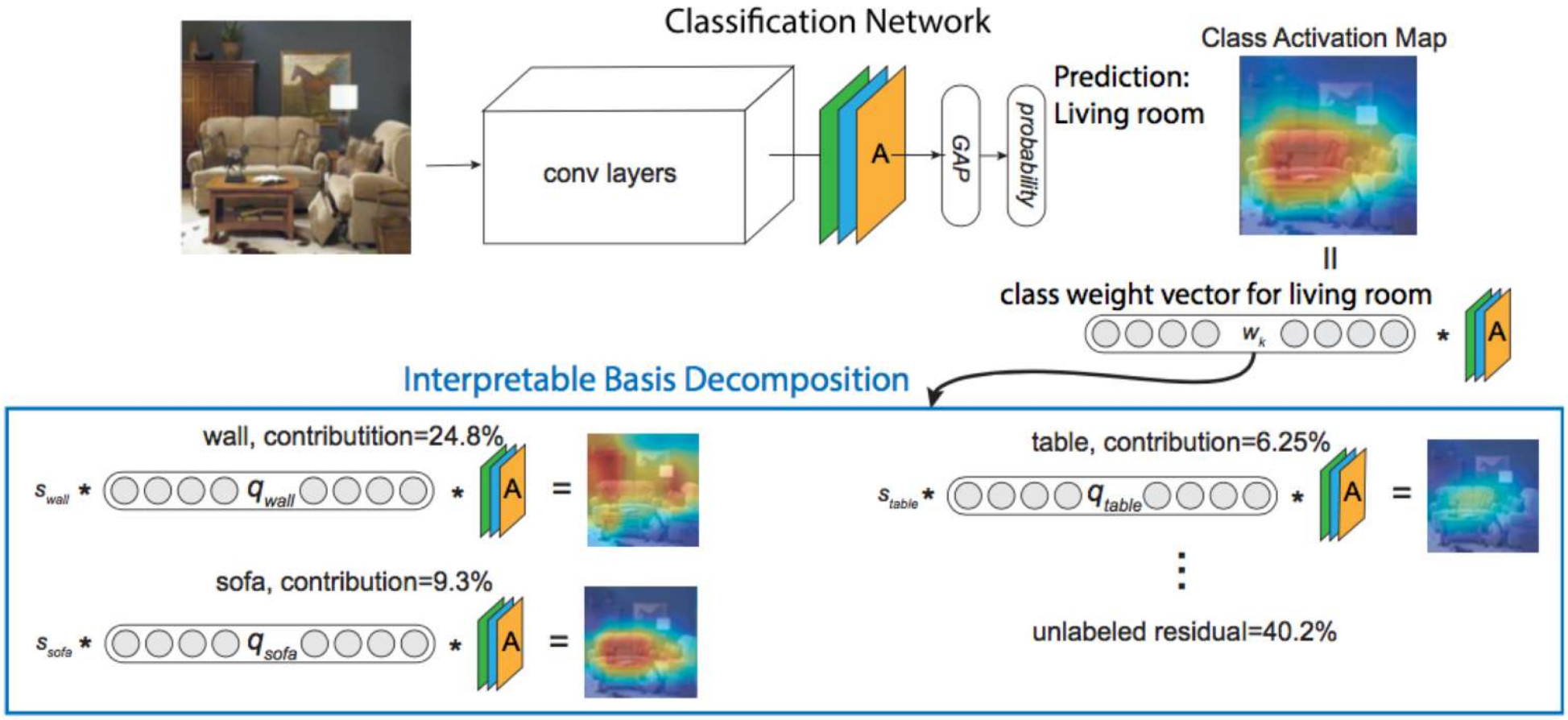


[1] Ramprasaath et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization”, ICCV 2017.

From attribution to interpretation

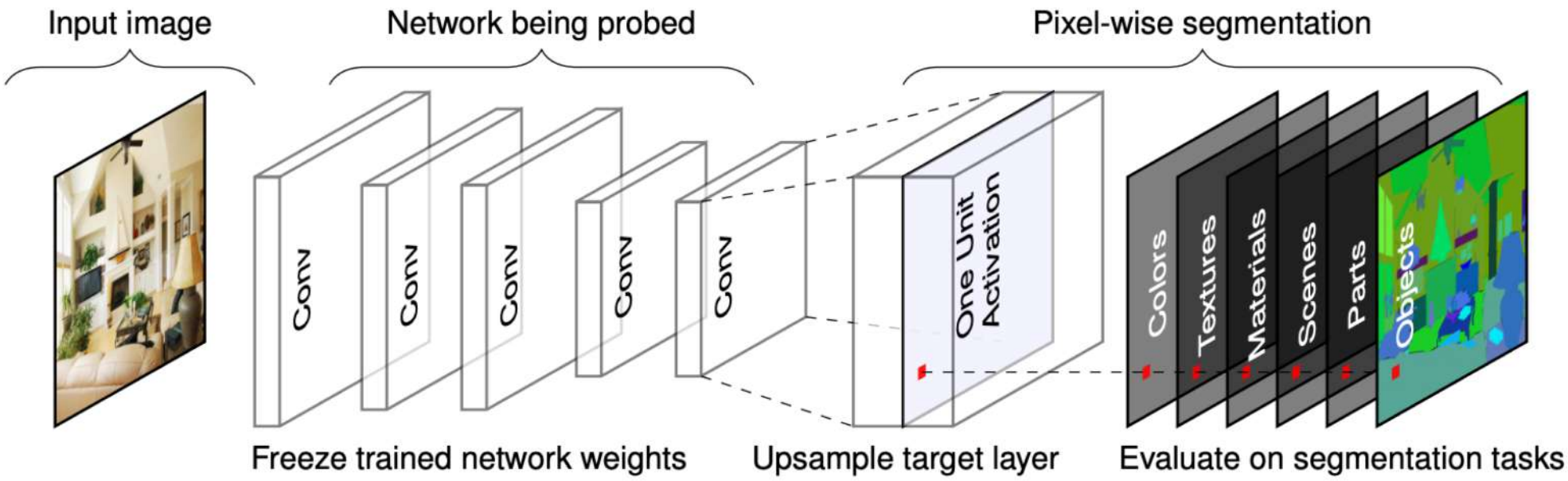


IBD



[1] Bolei et al, “Interpretable basis decomposition for visual explanation”, ECCV 2018

Train Concept Vectors



[1] Zhou et al., Network Dissection. PAMI 2018.

IBD Framework

Weight Vector: $w_k \in \mathbb{R}^D$

Interpretable Basis: $q_{c_i} \in \mathbb{R}^D \quad Q_c = [q_{c_1}|\cdots|q_{c_n}]$

Coefficient for the Interpretable Basis: $s_{c_i} \in \mathbb{R}$

$s = [s_1|\cdots|s_n]$

Residual: $r \in \mathbb{R}^D$

TASK:

$$\text{Find } s_{c_i} \text{ to minimize } \|r\| \text{ where } w_k = s_{c_1} q_{c_1} + \cdots + s_{c_n} q_{c_n} + r = Q_c s + r$$

Greedy Algorithm

Given that we have already chosen a set of columns

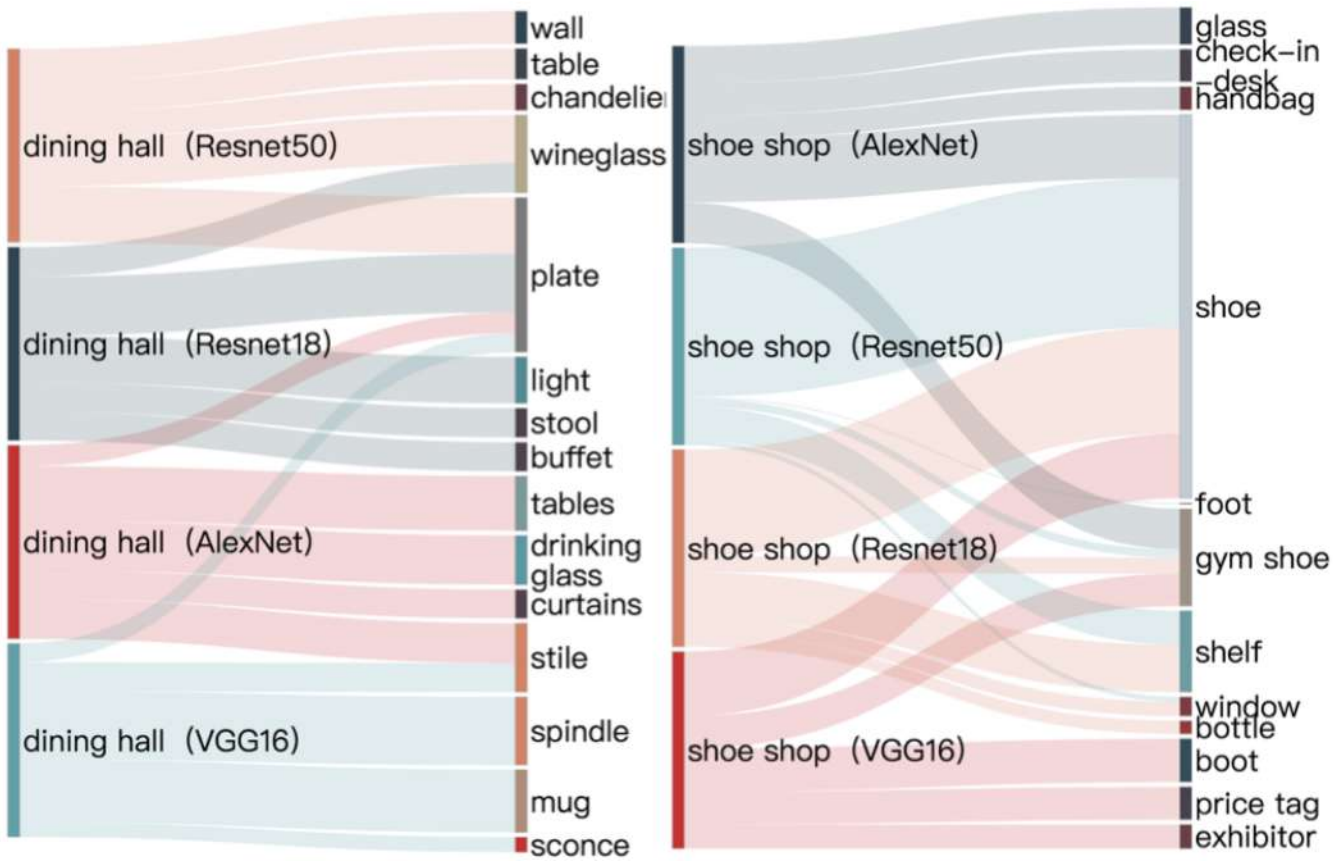
$$Q_{c_n} = [q_{c_1} | \cdots | q_{c_n}]$$

Find the (n + 1)th concept basis

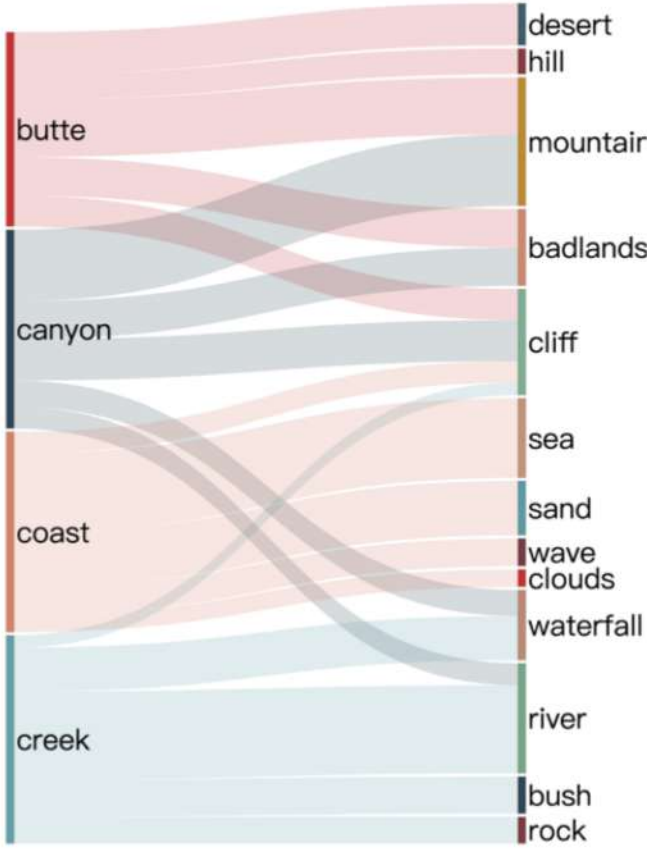
$$\underset{q_c i \in Q_c}{\operatorname{argmin}} \min_{s, s_i > 0} \| w_k - [Q_c \mid q_{c_i}] [s \mid s_i] \|$$

Explaining Classification Decision Boundaries

Comparing different concepts that different networks (Resnet18, Resnet50, AlexNet, VGG16) utilize to make predictions

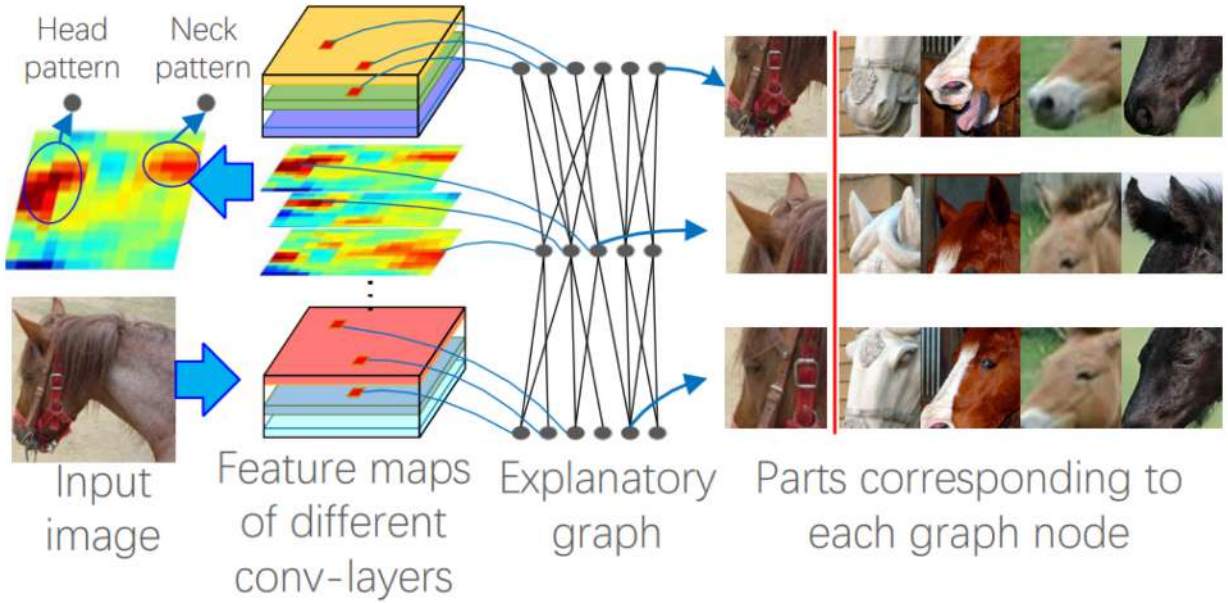


Comparing different concepts that Resnet18 utilizes to make different predictions.

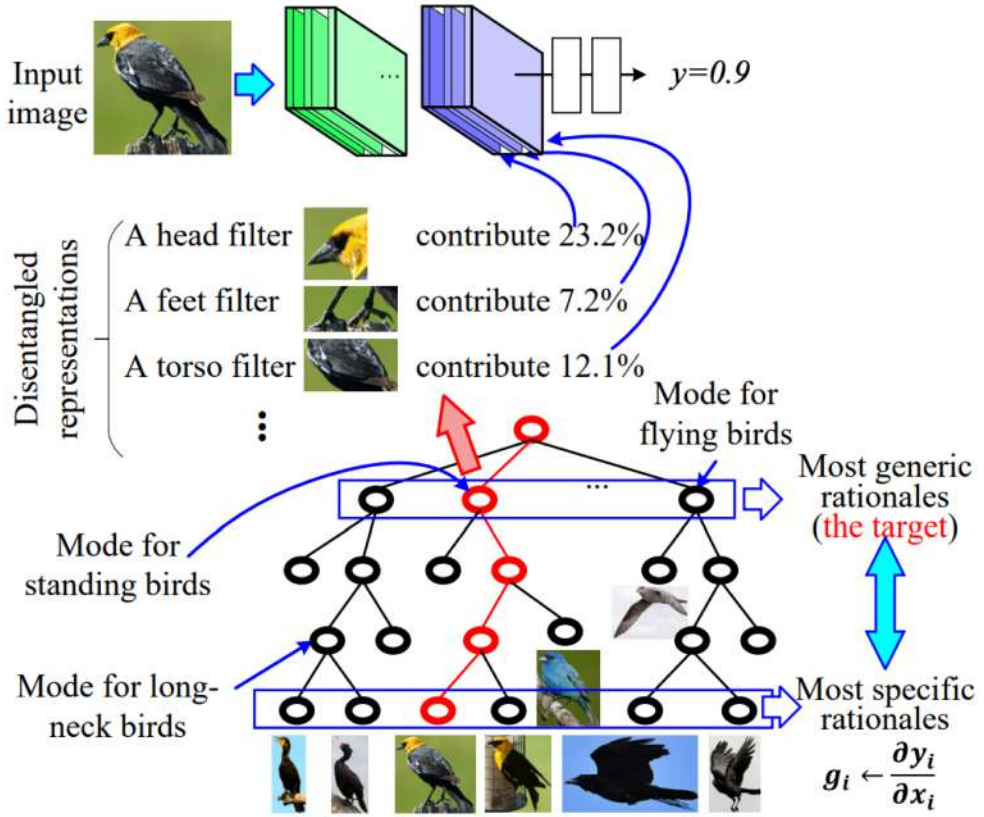


More Complicated Interpretation Methods

Interpret as Explanation Graph



Interpret as Decision Tree



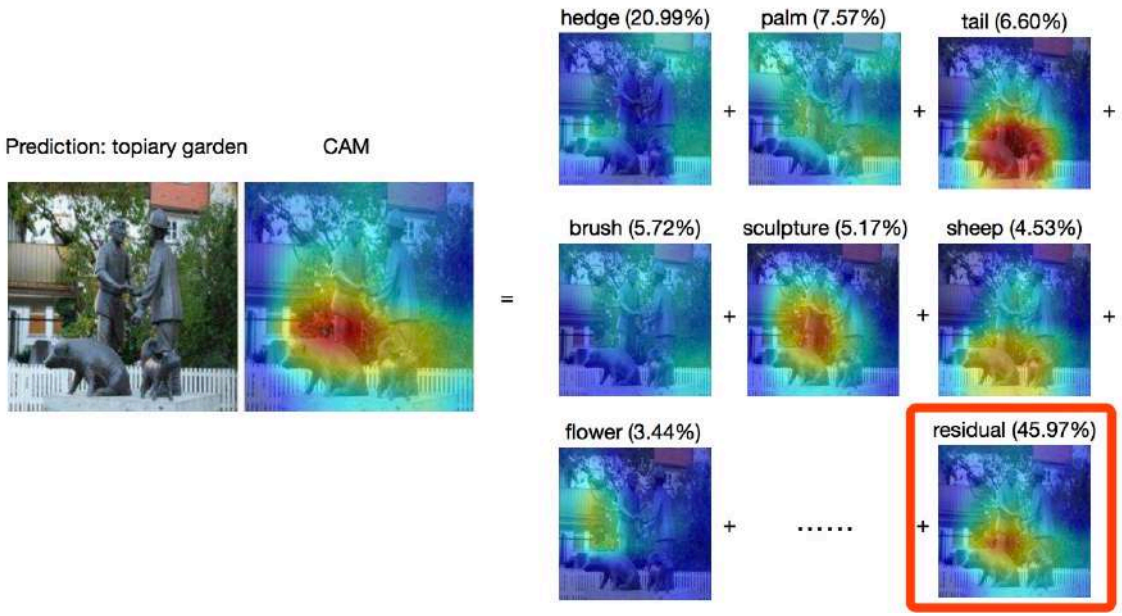
[1] Quanshi Zhang, Yu Yang, Ying Nian Wu, and Song-Chun Zhu. Interpreting cnns via decision trees. arXiv:1802.00121, 2018.
[2] Q. Zhang, R. Cao, F. Shi, Y.N. Wu, and S.-C. Zhu. Interpreting cnn knowledge via an explanatory graph. In AAAI, 2018.

Outline

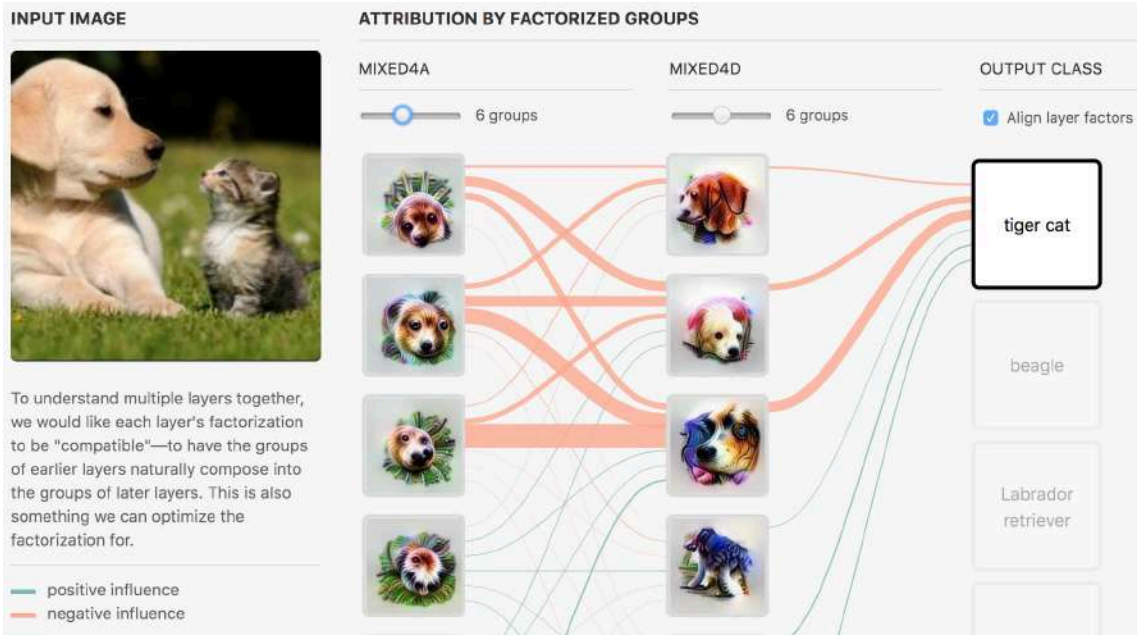
1. Visualization Methods
2. Attribution Methods
3. **Interpretable System**
4. Interpretation for GAN

What are the problems in interpretation methods?

IBD Framework



Building Block for Interpretability



[1] Bolei et al, “Interpretable basis decomposition for visual explanation”, ECCV 2018
[2] https://distill.pub/2018/building-blocks/

What are the problems in interpretation methods?

(ii) Explainable ML methods provide explanations that are not faithful to what the original model computes.

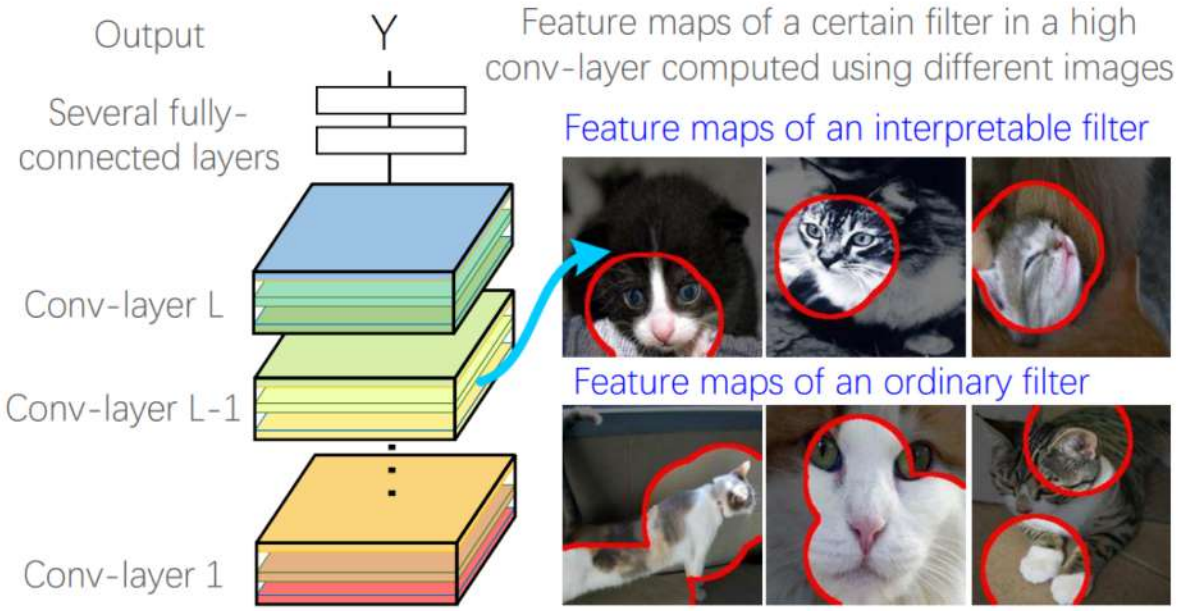
Explanations must be wrong. They cannot have perfect fidelity with respect to the original model. If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation. (In other words, this is a case where the original model would be interpretable.) This leads to the danger that any explanation method for a black box model can be an inaccurate representation of the original model in parts of the feature space. [See also for instance, 23, among others.]

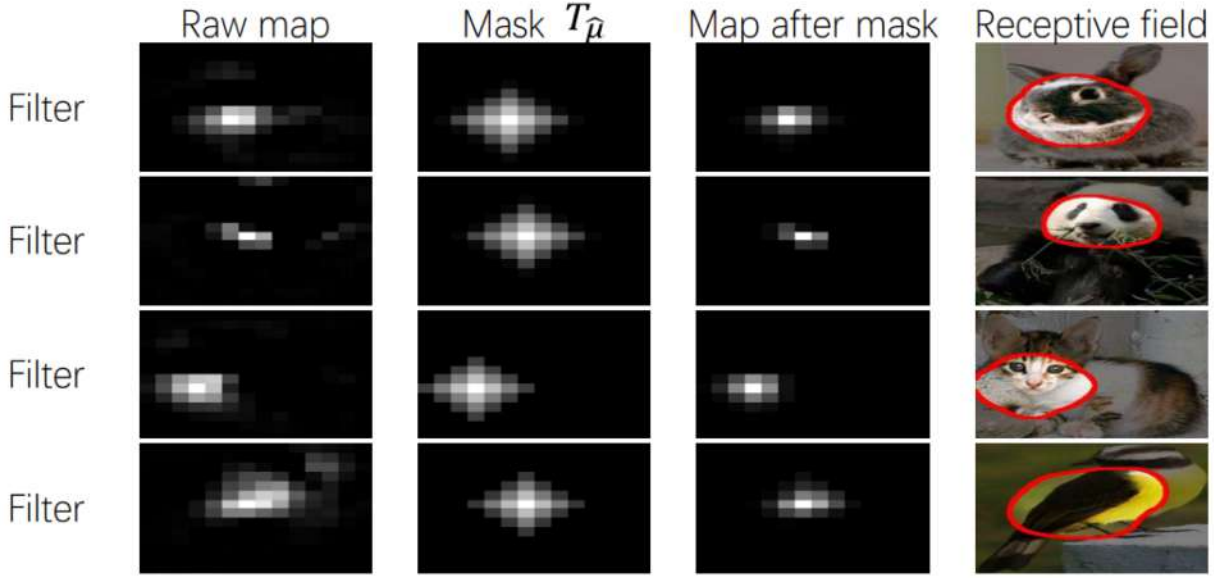
An inaccurate (low-fidelity) explanation model limits trust in the explanation, and by extension, trust in the black box that it is trying to explain. An explainable model that has a 90% agreement with the original model indeed explains the original model most of the time. However, an explanation model that is correct 90% of the time is wrong 10% of the time. If a tenth of the explanations are incorrect, one cannot trust the explanations, and thus one cannot trust the original black box. If we cannot know for certain whether our explanation is correct, we cannot know whether to trust either the explanation or the original model.

[1] Cynthia Rudin, “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”, arxiv:1811.10154.

How to optimize a
network
that is
by design,
interpretable?

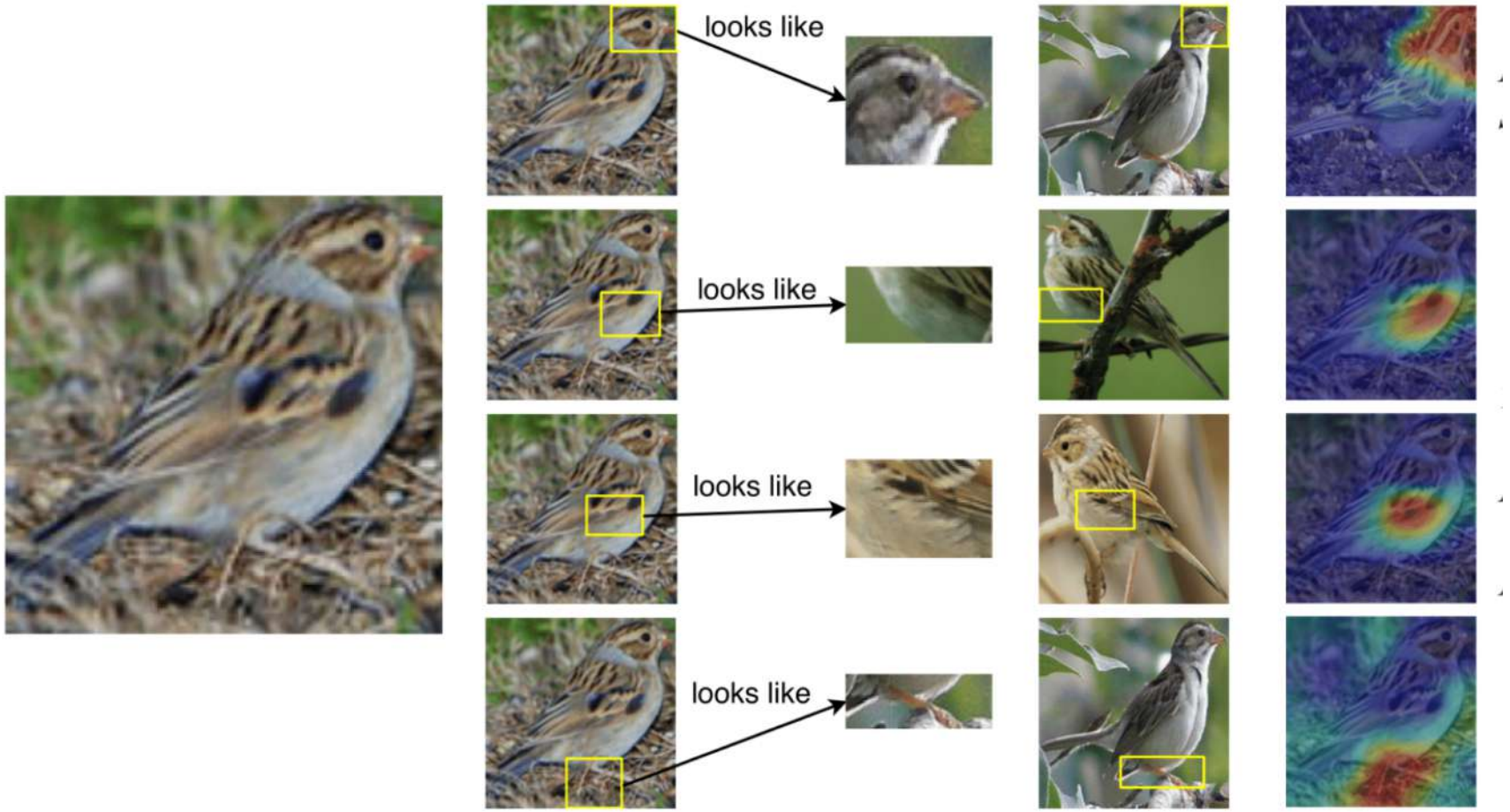
We can impose constraints on filters directly





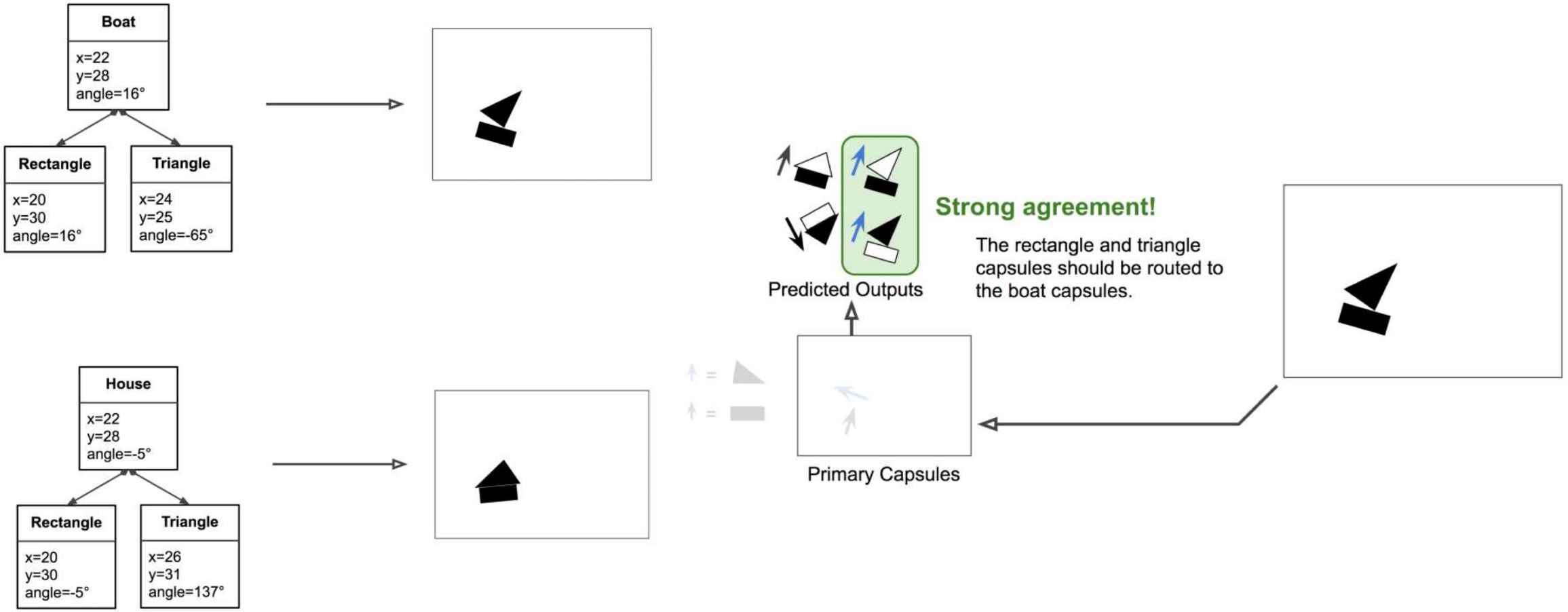
[1] Quanshi Zhang, “Interpretable Convolutional Neural Networks”, CVPR 2018.

We can train some “prototype” for inference



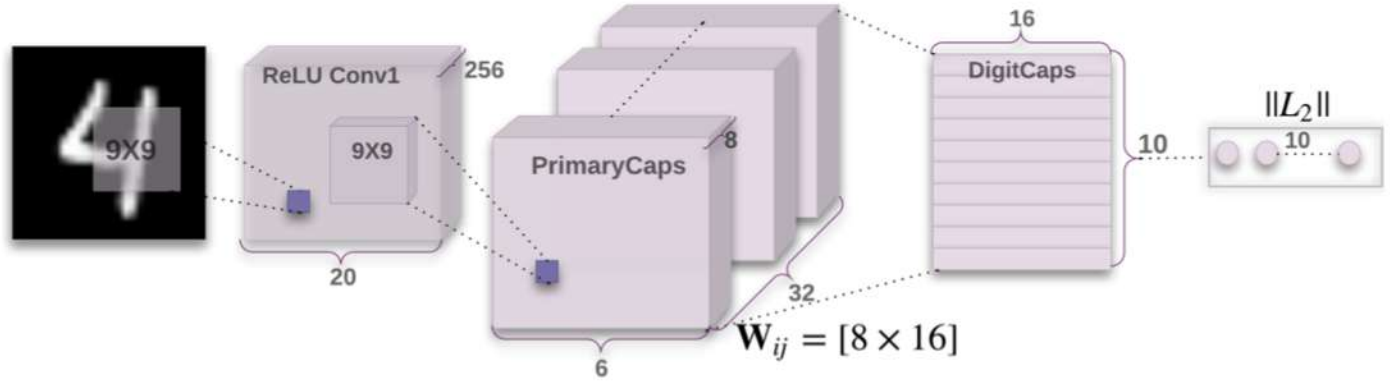
[1] Chaofan et al. “This Looks Like That: Deep Learning for Interpretable Image Recognition” NeurIPS spotlight.

Capsule Network: Capsule Concepts and Assemble



[1] https://www.youtube.com/watch?v=pPN8d0E3900

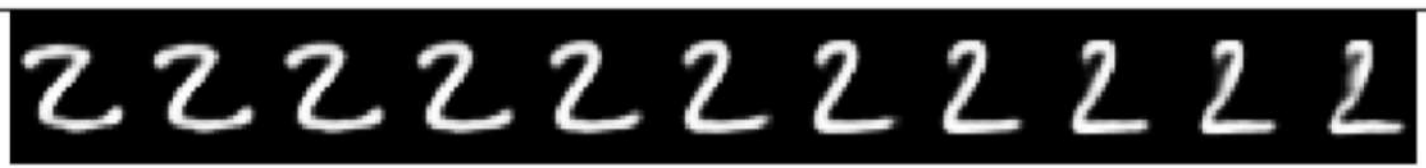
DYNAMIC CONNECTION



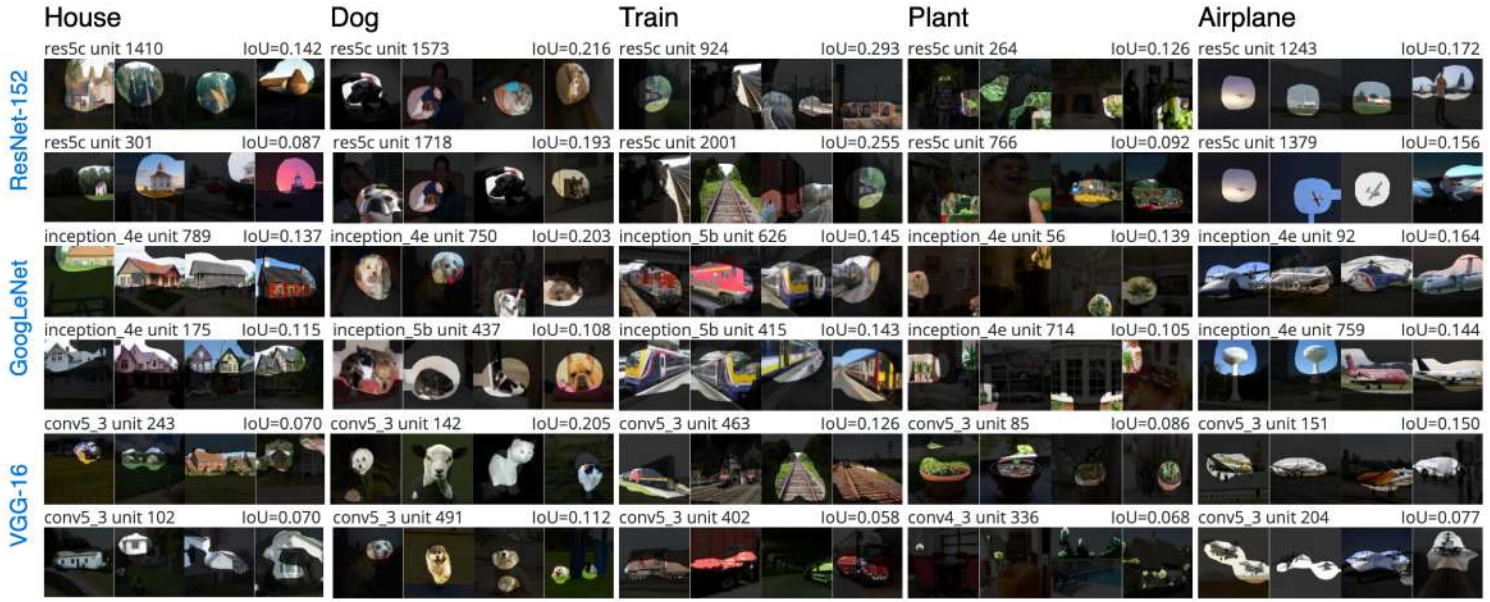
Procedure 1 Routing algorithm.

procedure ROUTING($\hat{\boldsymbol{u}}_{j|i}, r, l$)
  for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow 0$.
  for r iterations **do**
    for all capsule i in layer l: $\mathbf{c}_i \leftarrow$ `softmax`$(\mathbf{b}_i)$ ▷ `softmax` computes Eq. 3
    for all capsule j in layer $(l+1)$: $\mathbf{s}_j \leftarrow \sum_i c_{ij} \hat{\mathbf{u}}_{j|i}$
    for all capsule j in layer $(l+1)$: $\mathbf{v}_j \leftarrow$ `squash`$(\mathbf{s}_j)$ ▷ `squash` computes Eq. 1
    for all capsule i in layer l and capsule j in layer $(l+1)$: $b_{ij} \leftarrow b_{ij} + \hat{\mathbf{u}}_{j|i}.\mathbf{v}_j$
  return $\mathbf{v}_j$

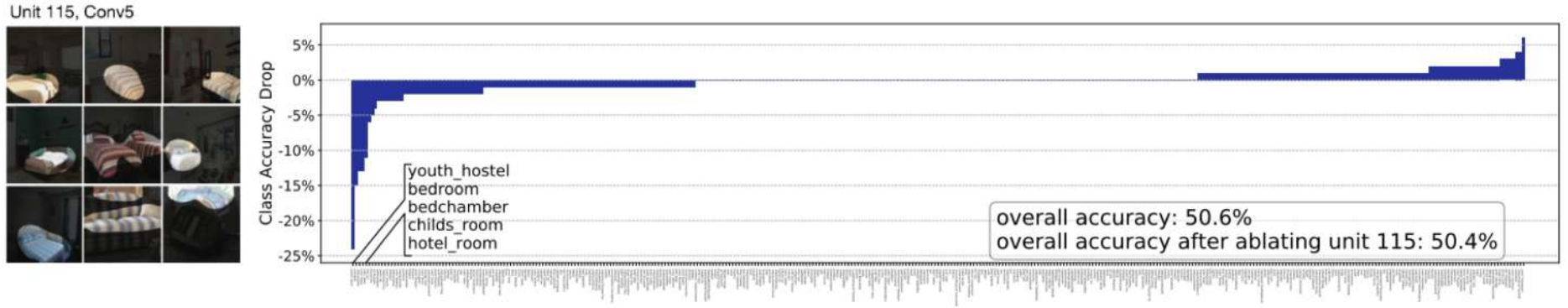
[1] Sara et al. “Dynamic Routing Between Capsules”. NeurIPS 2018.

Scale and thickness	
Localized part	
Stroke thickness	
Localized skew	
Width and translation	
Localized part	

[1] Sara et al. “Dynamic Routing Between Capsules”. NeurIPS 2018.

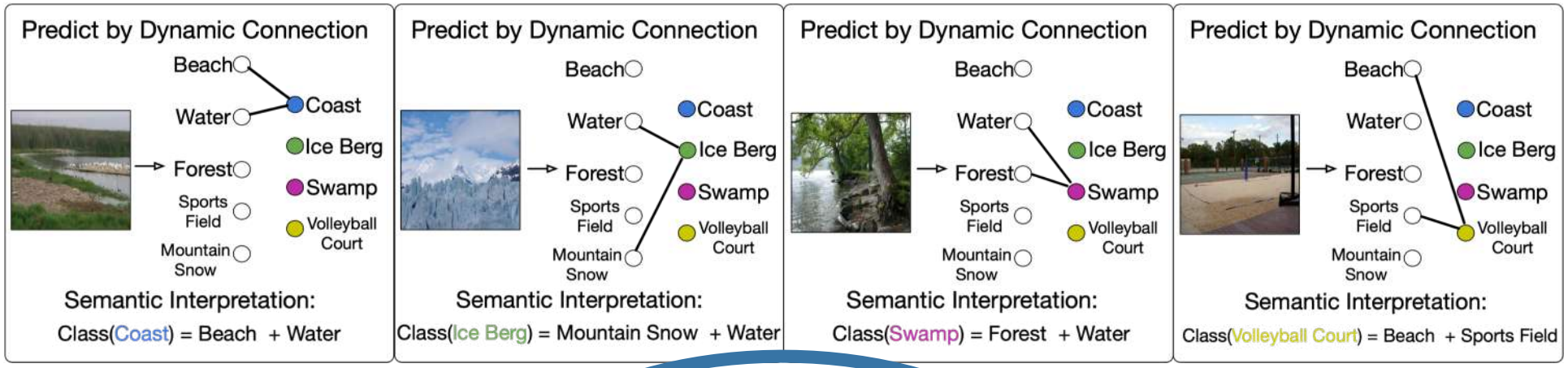


[David. B et al., CVPR, 2017]



[Bolei. Z et al., arxiv, 2018]

IMPLICATION WITH DYNAMIC CONNECTION



Sparsity

It reduces the parameters to be analyzed.

Simulatability

Human is able to simulate its decision-making process.

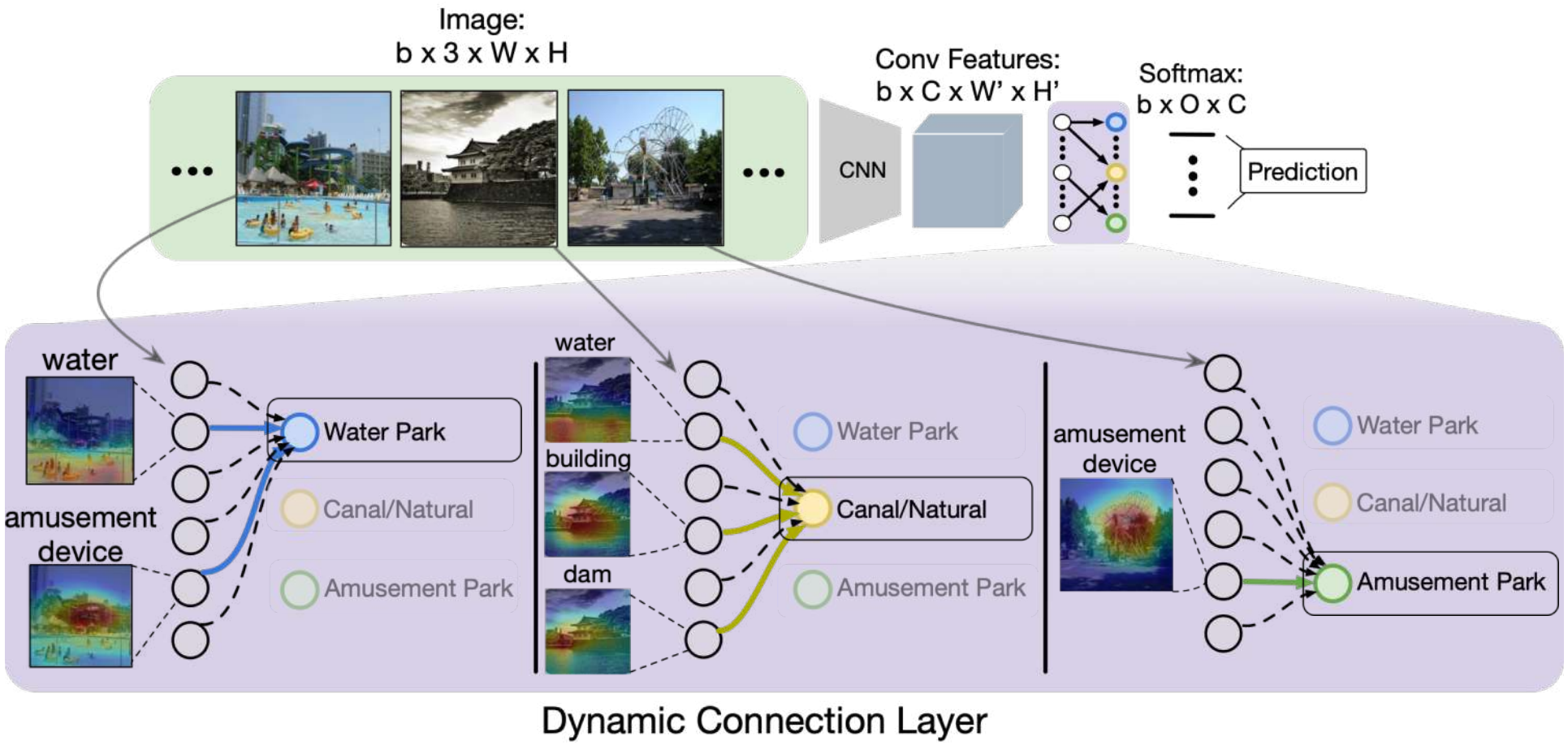
Modularity

The meaningful units can be interpreted independently

Interpretability

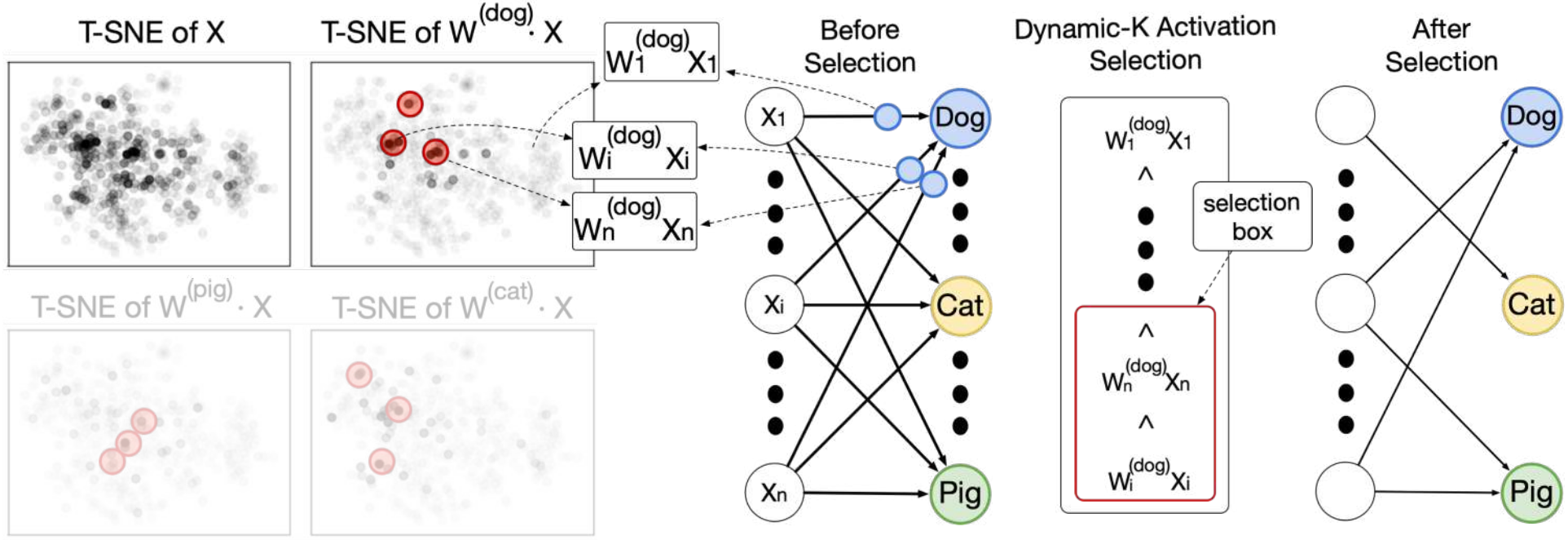
[1] Yiyou et al. "Adaptive Activation Thresholding: Dynamic Routing Type Behavior for Interpretability in Convolutional Neural Networks". ICCV 2019.

MAIN INTUITION



[1] Yiyou et al. "Adaptive Activation Thresholding: Dynamic Routing Type Behavior for Interpretability in Convolutional Neural Networks". ICCV 2019.

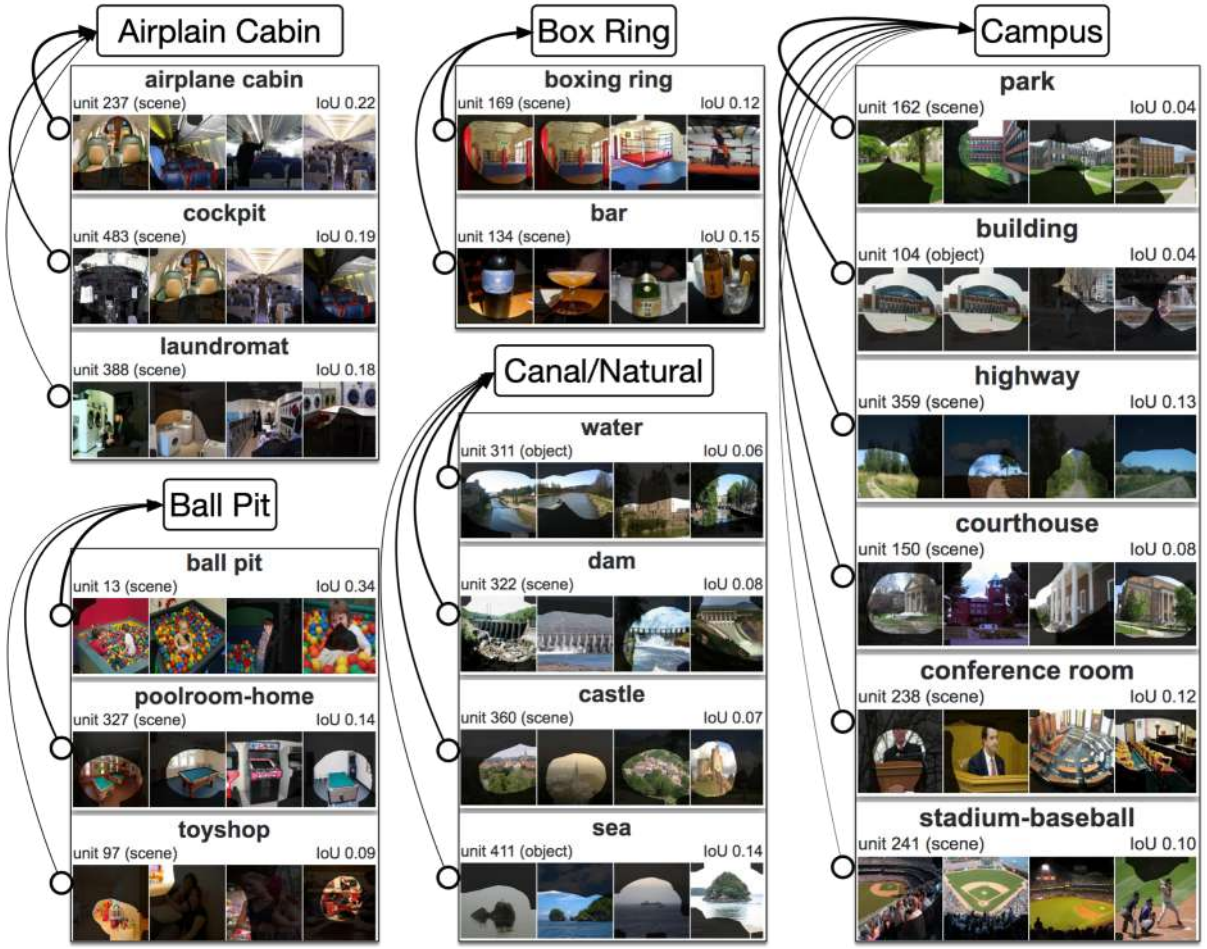
ROUTING ALGORITHM



[1] Yiyou et al. “Adaptive Activation Thresholding: Dynamic Routing Type Behavior for Interpretability in Convolutional Neural Networks”. ICCV 2019.

EXPERIMENT

CONCEPT COMPOSABILITY

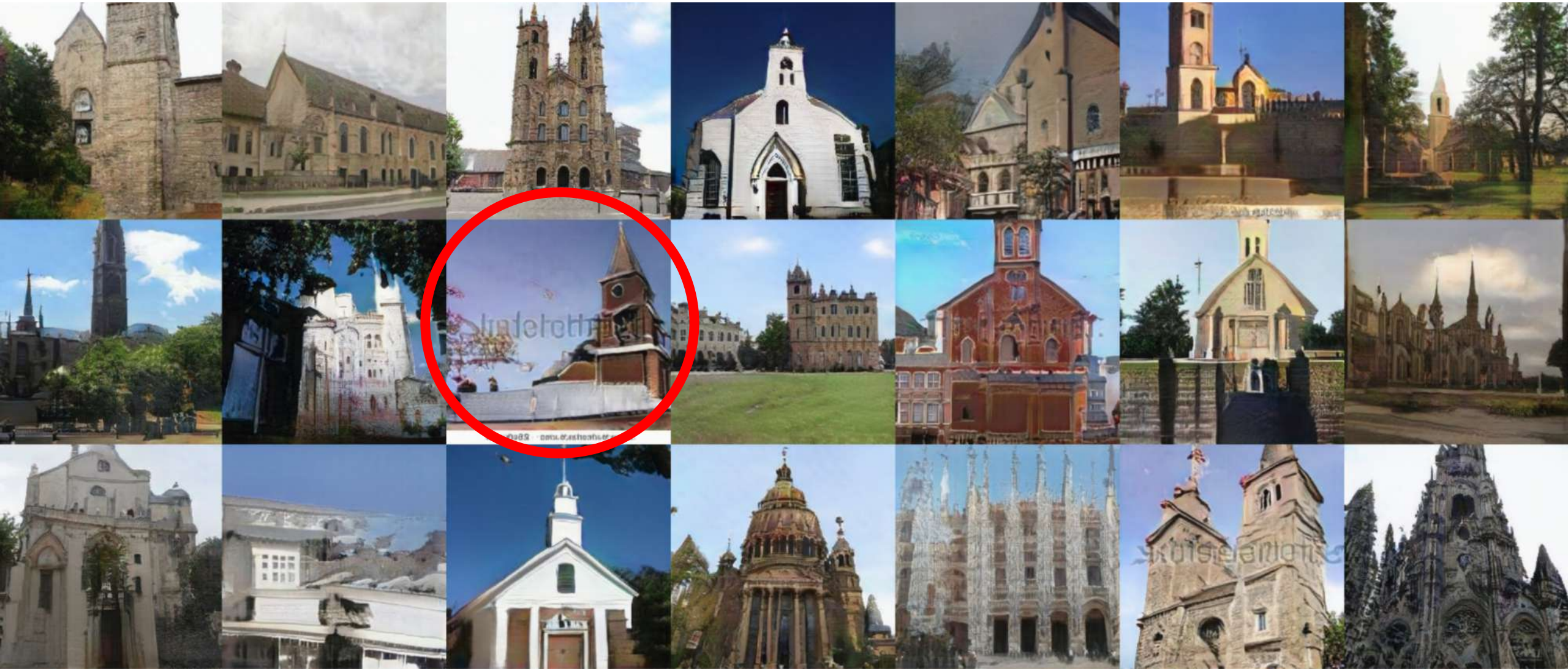


[1] Yiyou et al. “Adaptive Activation Thresholding: Dynamic Routing Type Behavior for Interpretability in Convolutional Neural Networks”. ICCV 2019.

Outline

1. Visualization Methods
2. Attribution Methods
3. Interpretable System
4. **Interpretation for GAN**

ProgressiveGAN

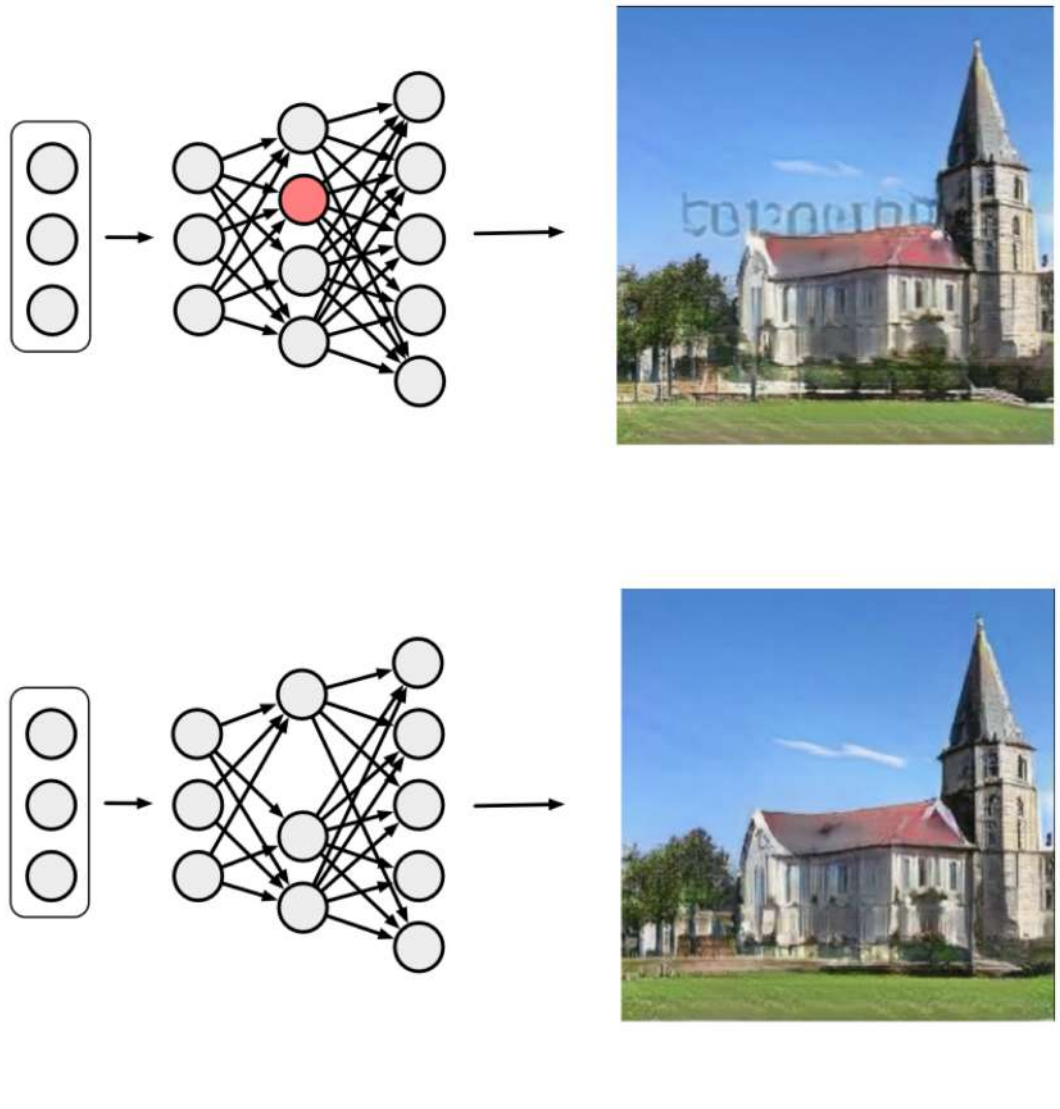


[1] http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_David_Talk.pdf

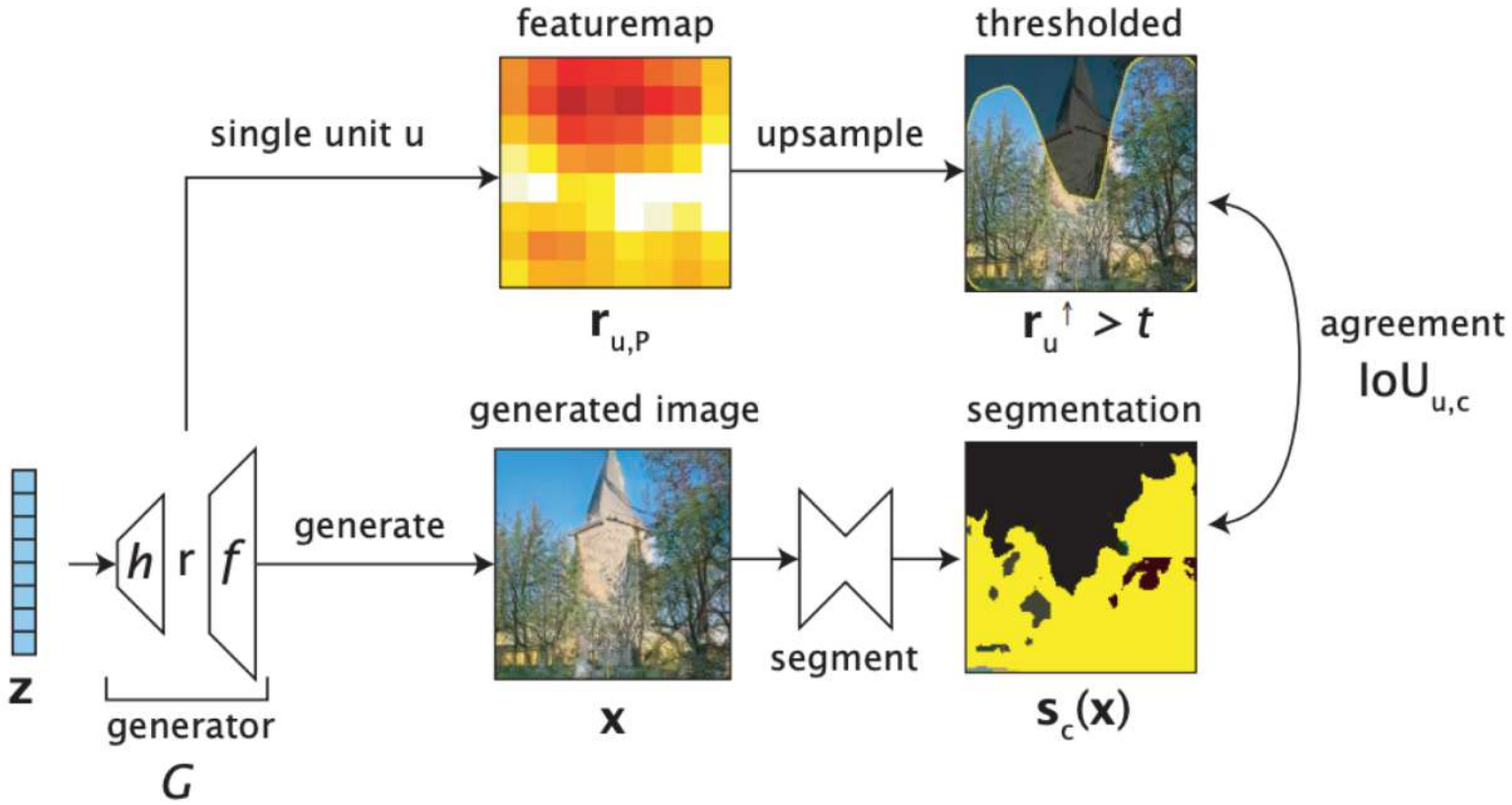
Layer 4, Neuron 201

Layer 4, Neuron 445

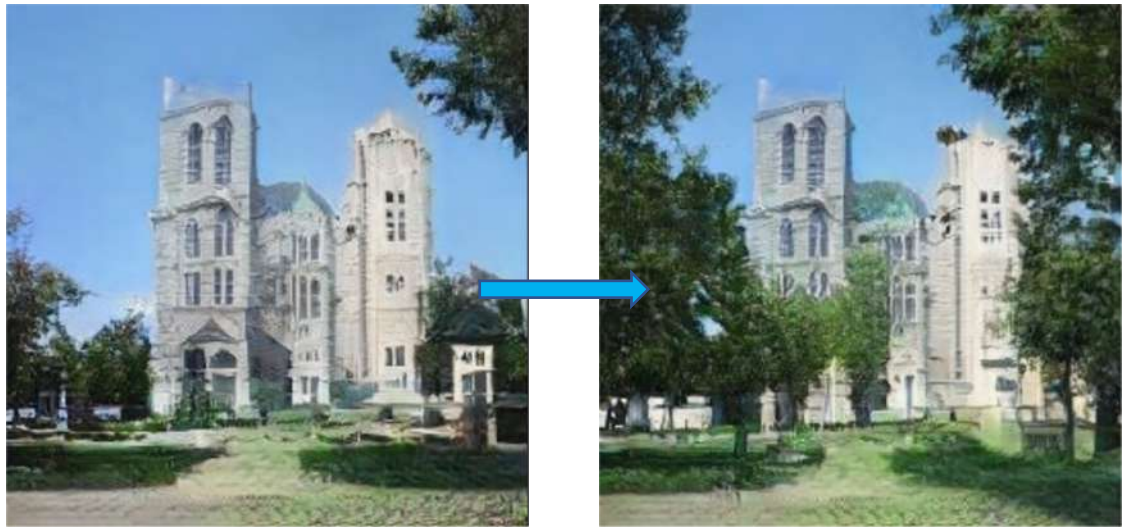
[1] http://xai.unist.ac.kr/static/img/event/ICCV_2019_VXAI_David_Talk.pdf

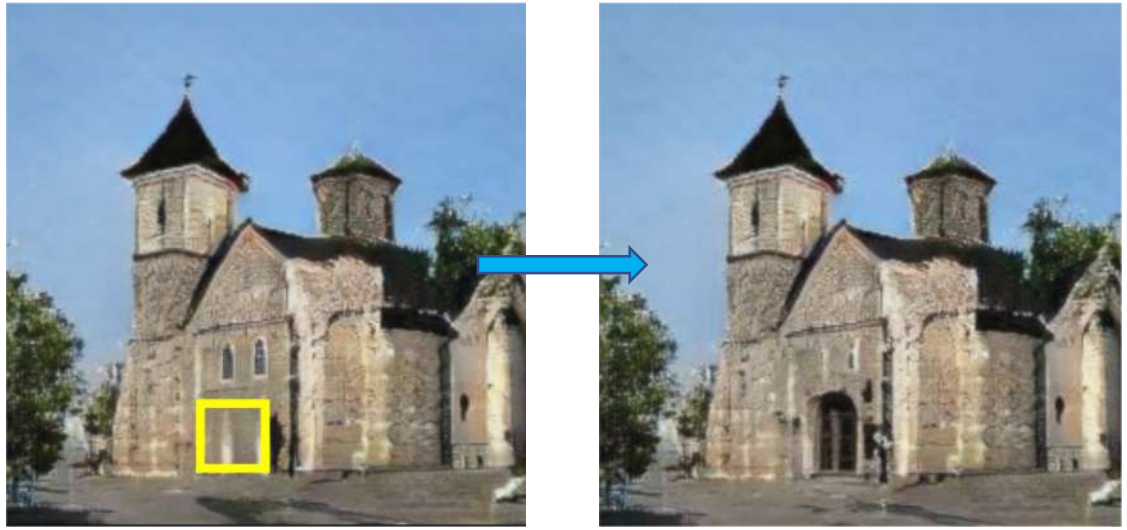


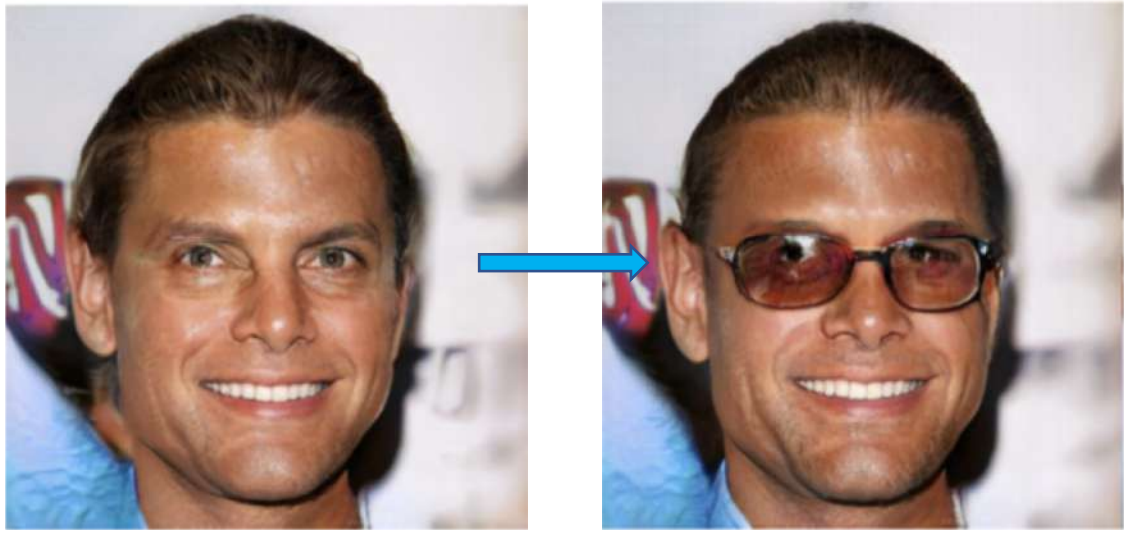
GAN Dissection



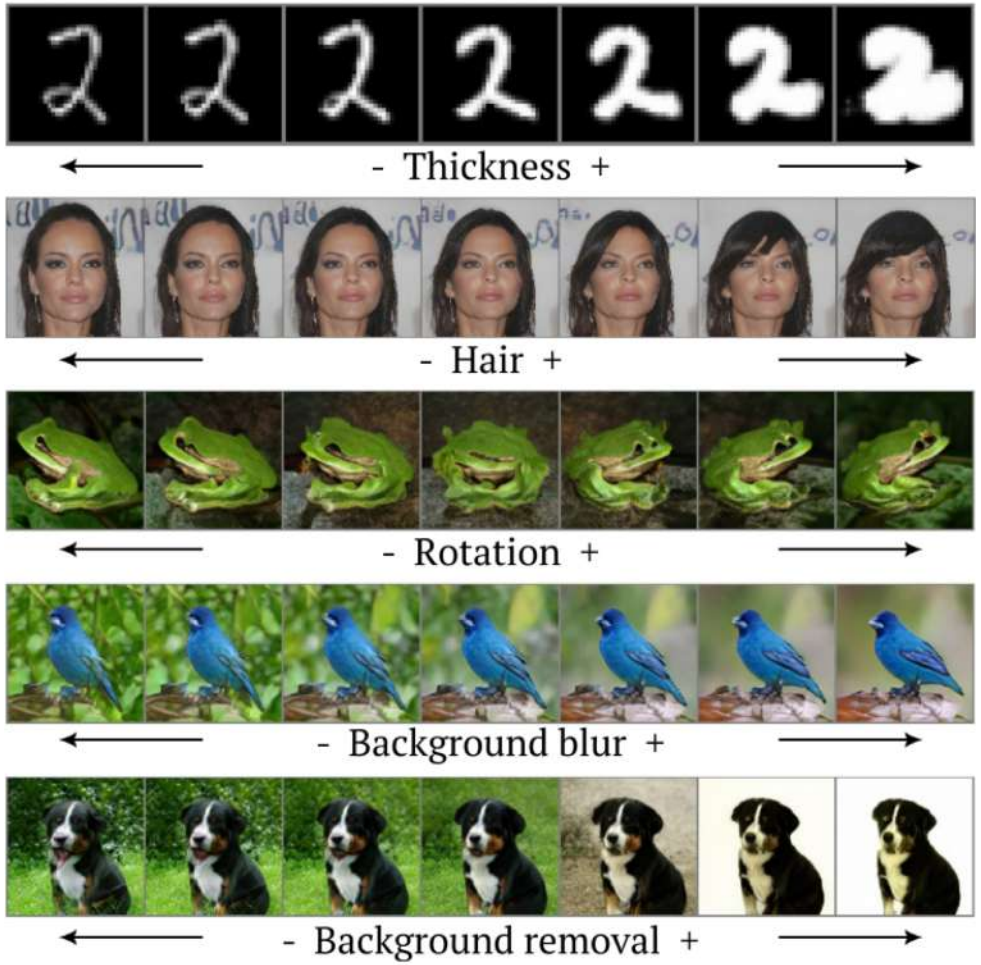
[1] David et al. “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks.” ICLR 2019.





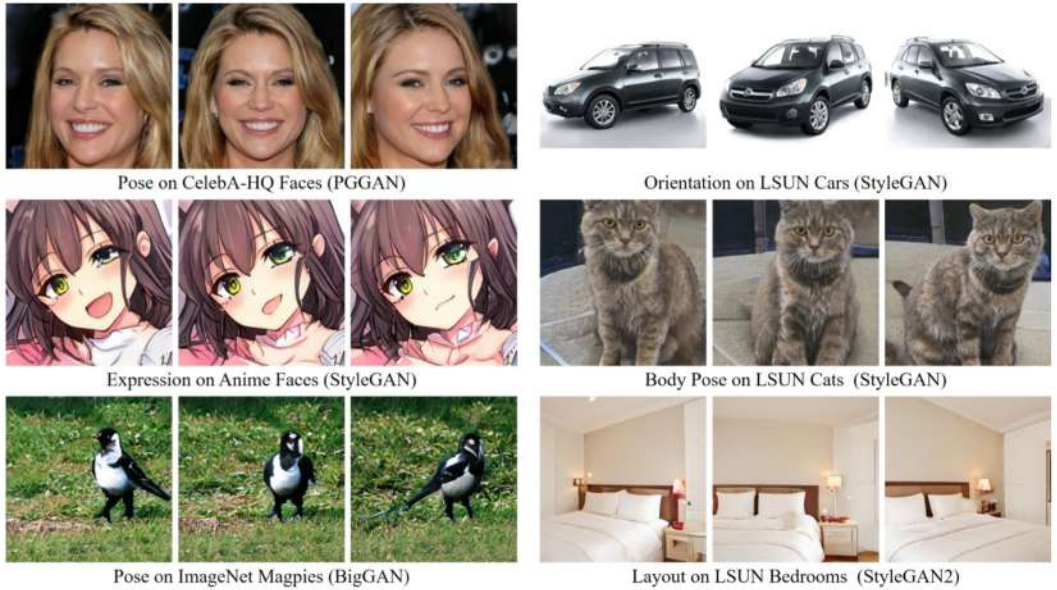


[1] David et al. “GAN Dissection: Visualizing and Understanding Generative Adversarial Networks.” ICLR 2019.



[Andrey et al. arxiv: 2002.03754]

[William et al. ECCV 2020]



[Yujun et al. arXiv:2007.06600]

Thanks!

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON