Day 6 Lecture 1

Interpretability

Explainable AI (XAI)

Xavier Giro-i-Nieto

@DocXavi
xavier.giro@upc.edu

Associate Professor
Universitat Politècnica de Catalunya

telecos BCN

UPC

Versió extesa

Xavier Giró (en català)
[UPC AA2 2020]

Acknowledgements

Amaia Salvador
amaia.salvador@upc.edu

PhD 2019
Universitat Politècnica de Catalunya

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Eva Mohedano
eva.mohedano@insight-centre.org

PhD 2018
Insight-centre for Data Analytics
Dublin City University

DCU

Insight
Centre for Data Analytics

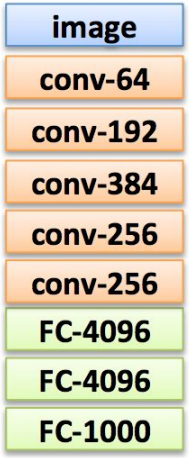
Videolectures

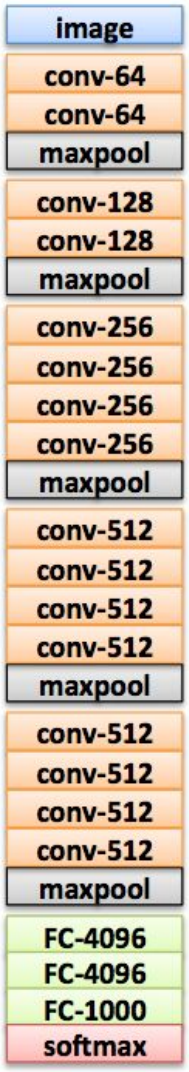
Amaia Salvador
[UPC DLCV 2016]

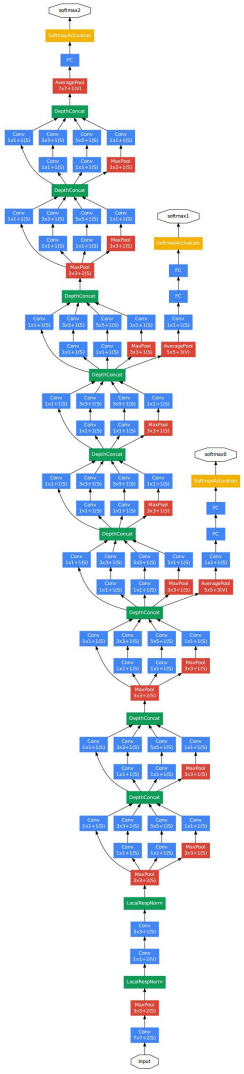
Eva Mohedano
[UPC DLCV 2018]

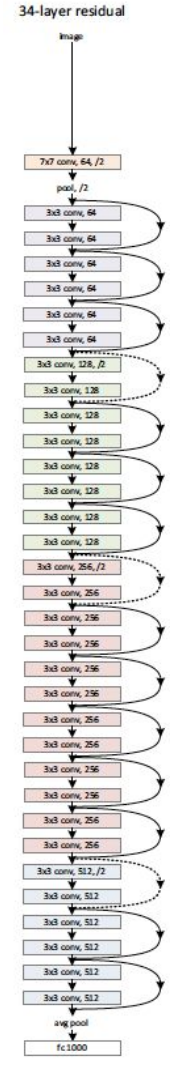
Laura Leal-Taixé
[UPC DLCV 2019]

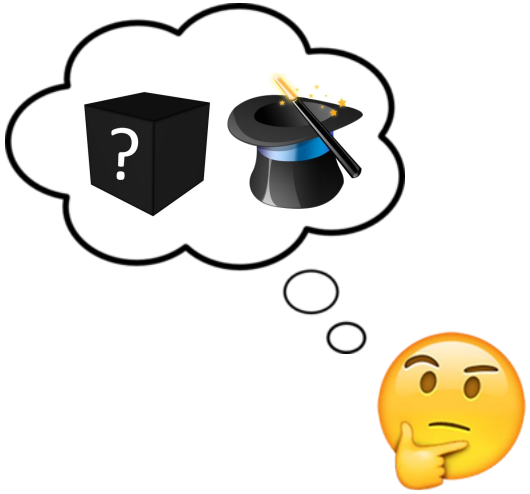
AlexNet



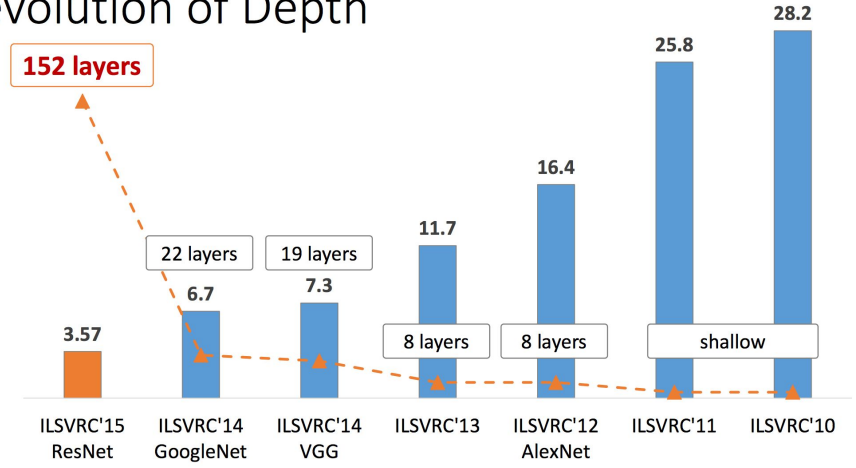






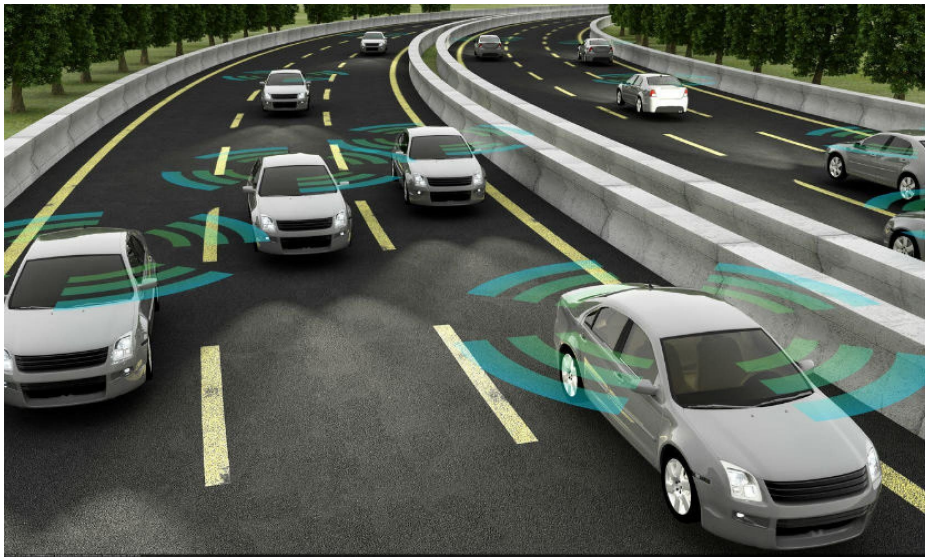


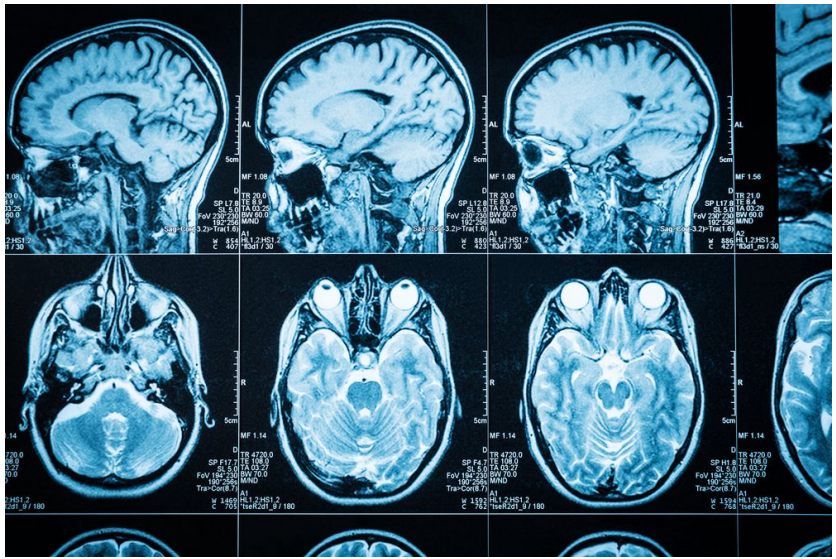
Revolution of Depth



Motivation

In many cases, an explanation for NN prediction is required.

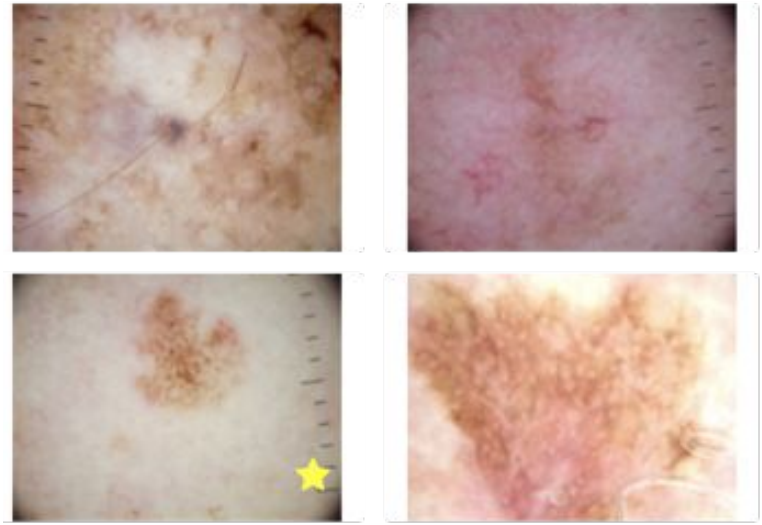




Motivation

"But, after its publication, the authors of the study noticed a bias in their algorithm — it was more likely to label an image as malignant cancer if there was a ruler in the image."

Nicole Wheeler, "Is the Media's Reluctance to Admit AI's Weaknesses Putting us at Risk?" (2019)



Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115.

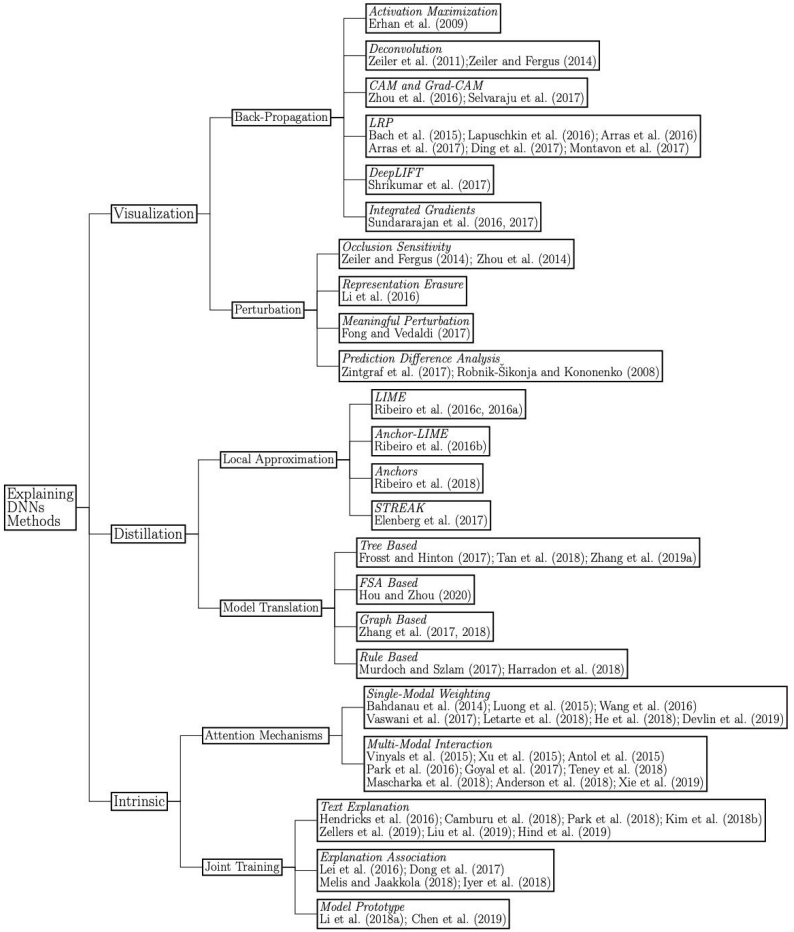
Motivation

Geoffrey Hinton
@geoffreyhinton

Suppose you have cancer and you have to choose between a black box AI surgeon that cannot explain how it works but has a 90% cure rate and a human surgeon with an 80% cure rate. Do you want the AI surgeon to be illegal?

Tradueix el tuit
9:37 p. m. · 20 de febr. de 2020 · Twitter Web App

Broad picture

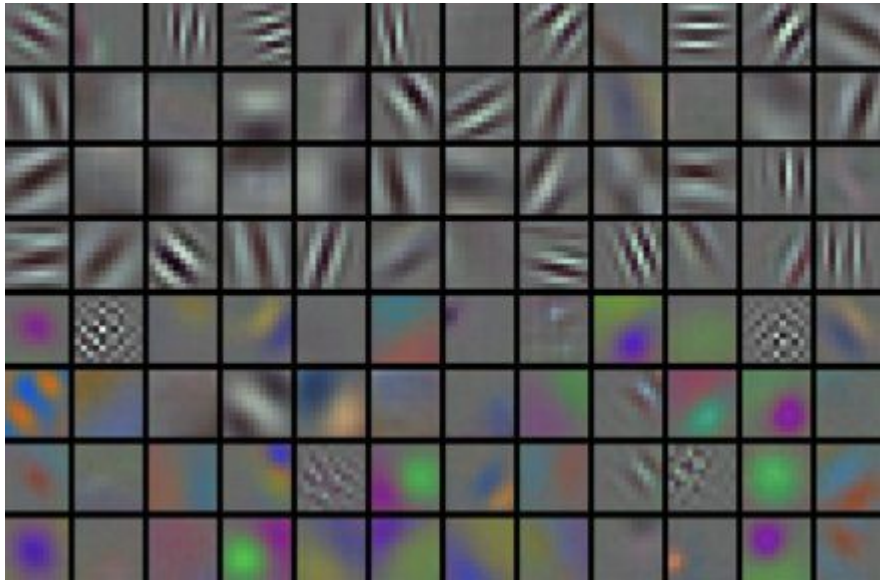


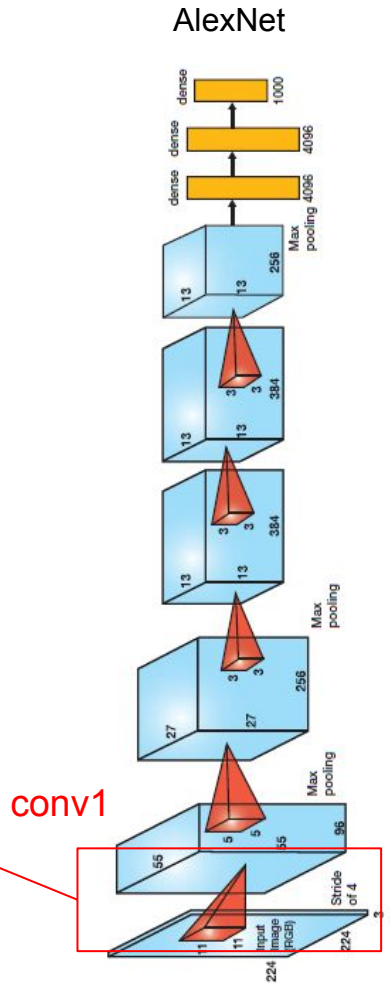
Ning Xie, Gabrielle Ras, Marcel van Gerven, Derek Doran, "Explainable Deep Learning: A Field Guide for the Uninitiated" arXiv 2020.

Interpretability

- **Visualization**
  - **Learned Weights**
  - Feature Maps
- Attribution
- Feature visualization

Visualization of Learned Weights

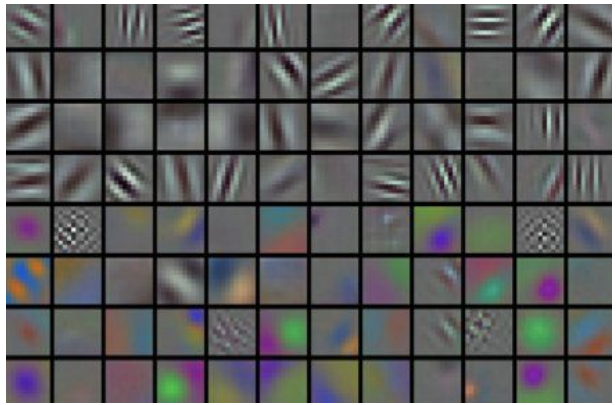


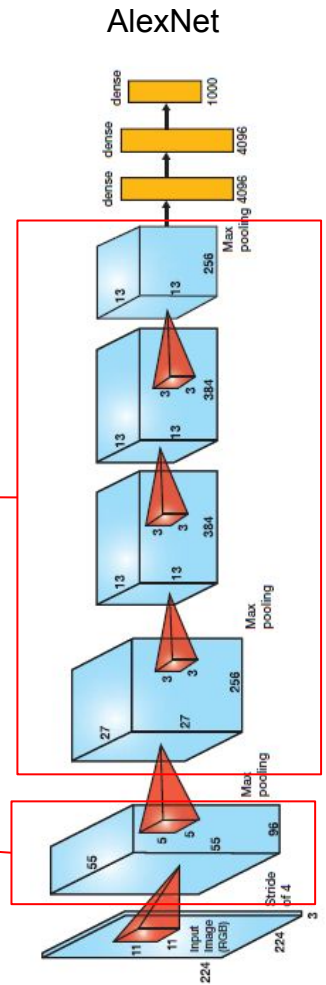


Visualization of Learned Weights

Can we observe the filters in other layers as RGB images ?

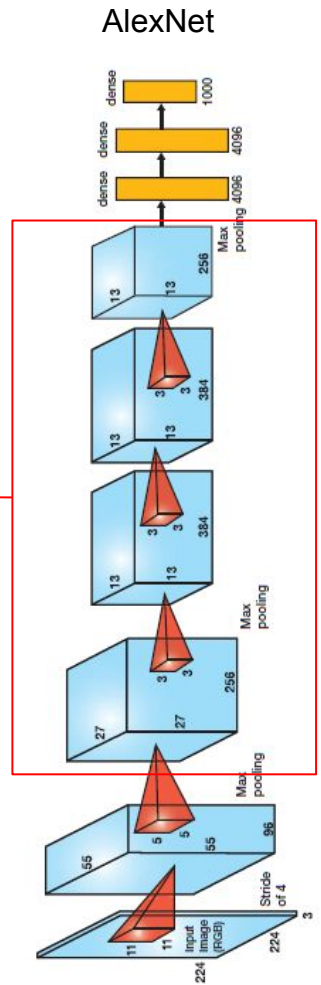
No, only convolutional filters from the first layer can be “visualized”, because their depth of 3 matches the input RGB channels.





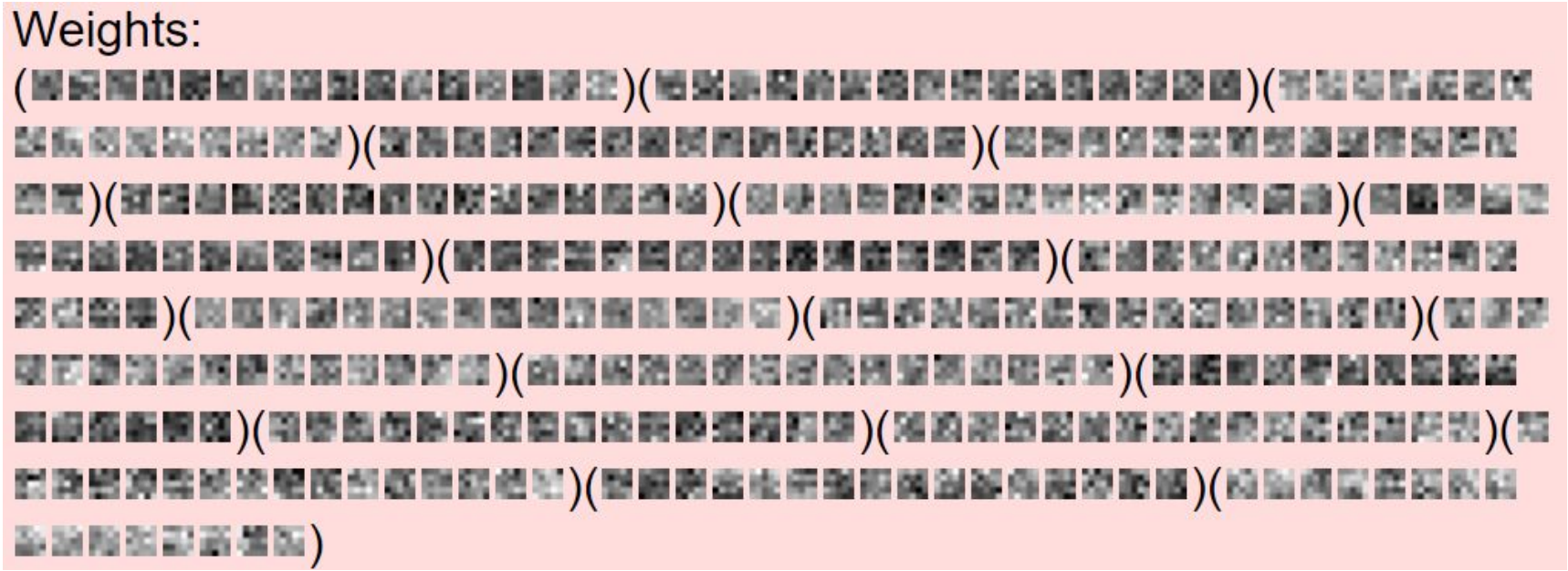
Visualization of Learned Weights

How can we observe the filters in deeper layers ?



Visualization of Learned Weights

Filters with depth larger than 3 can be visualized showing a gray-scale image of each depth, one next to the other.



2D Convolutional filters learned in Layer #2 from ConvnetJS

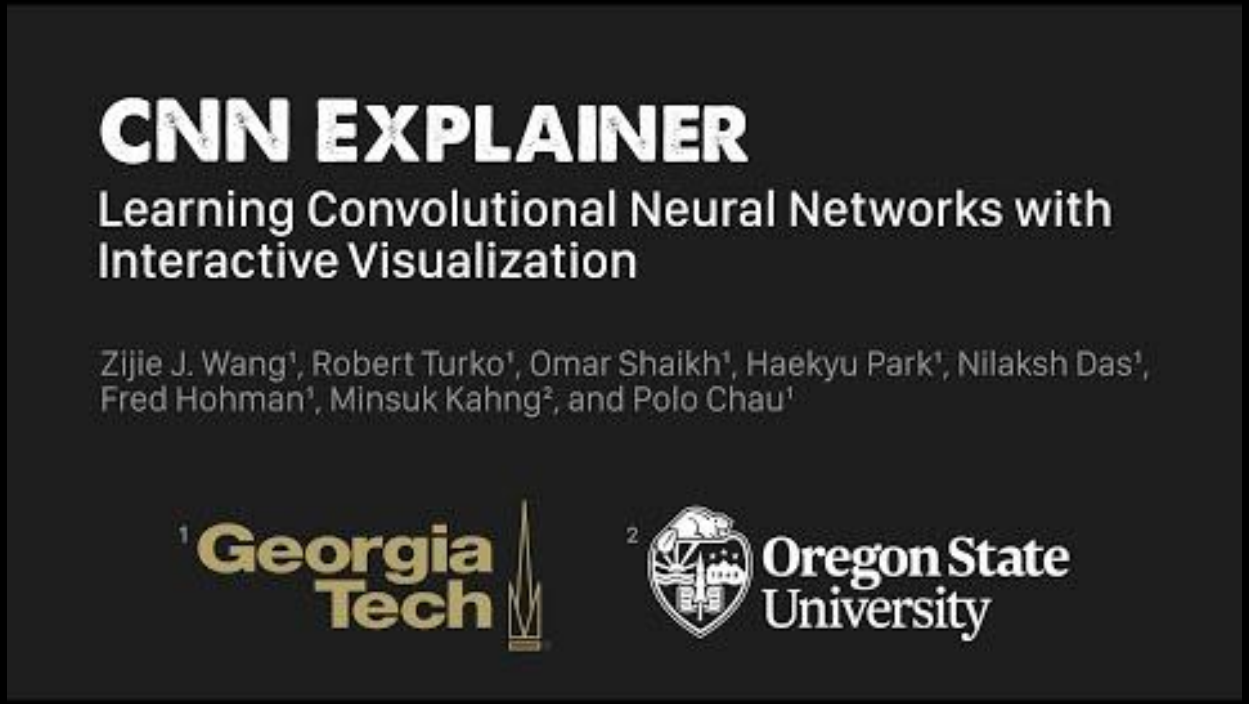
Visualization of Learned Weights

Demo: [Classify MNIST digits with a Convolutional Neural Network]

ConvNetJS

Deep Learning in your browser

"ConvNetJS is a Javascript library for training Deep Learning models (mainly Neural Networks) entirely in your browser. Open a tab and you're training. No software requirements, no compilers, no installations, no GPUs, no sweat."



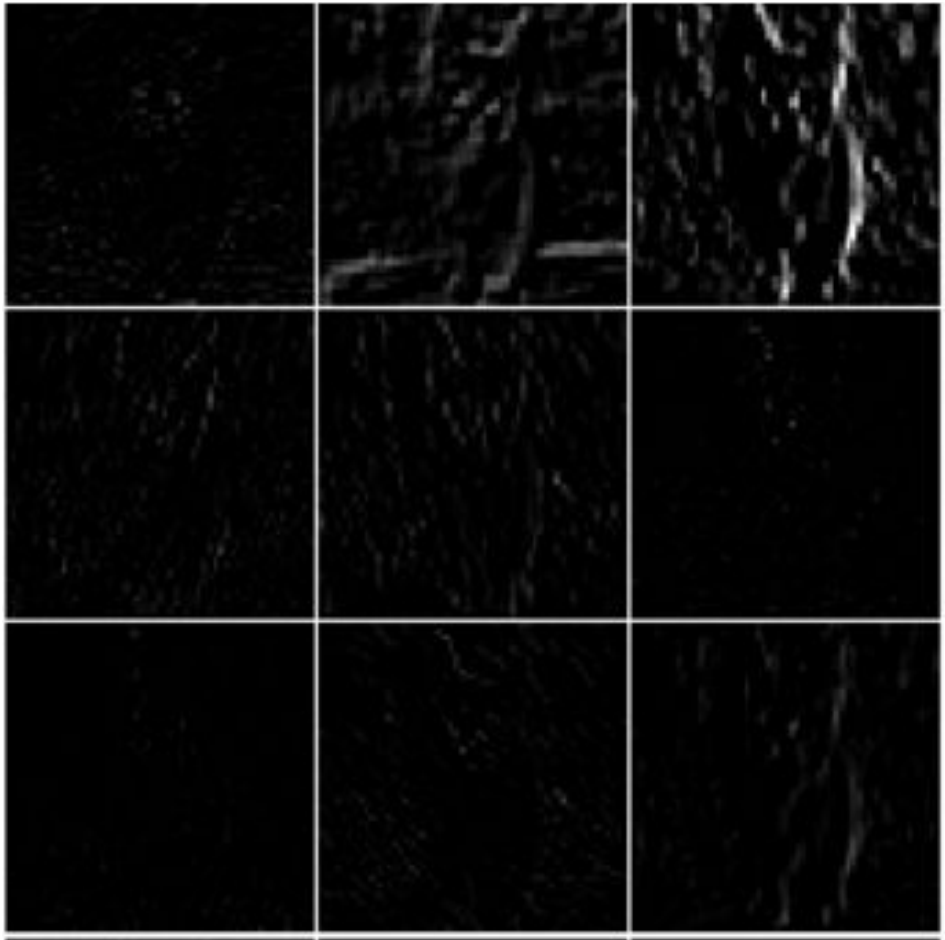
#CNNExplainer Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., ... & Chau, D. H. (2020). CNN Explainer: Learning Convolutional Neural Networks with Interactive Visualization. IEEE VIS 2020.

Interpretability

- **Visualization**
  - Learned Weights
  - **Feature Maps**
- Attribution
- Feature visualization

Visualization of Feature Activations (2D)

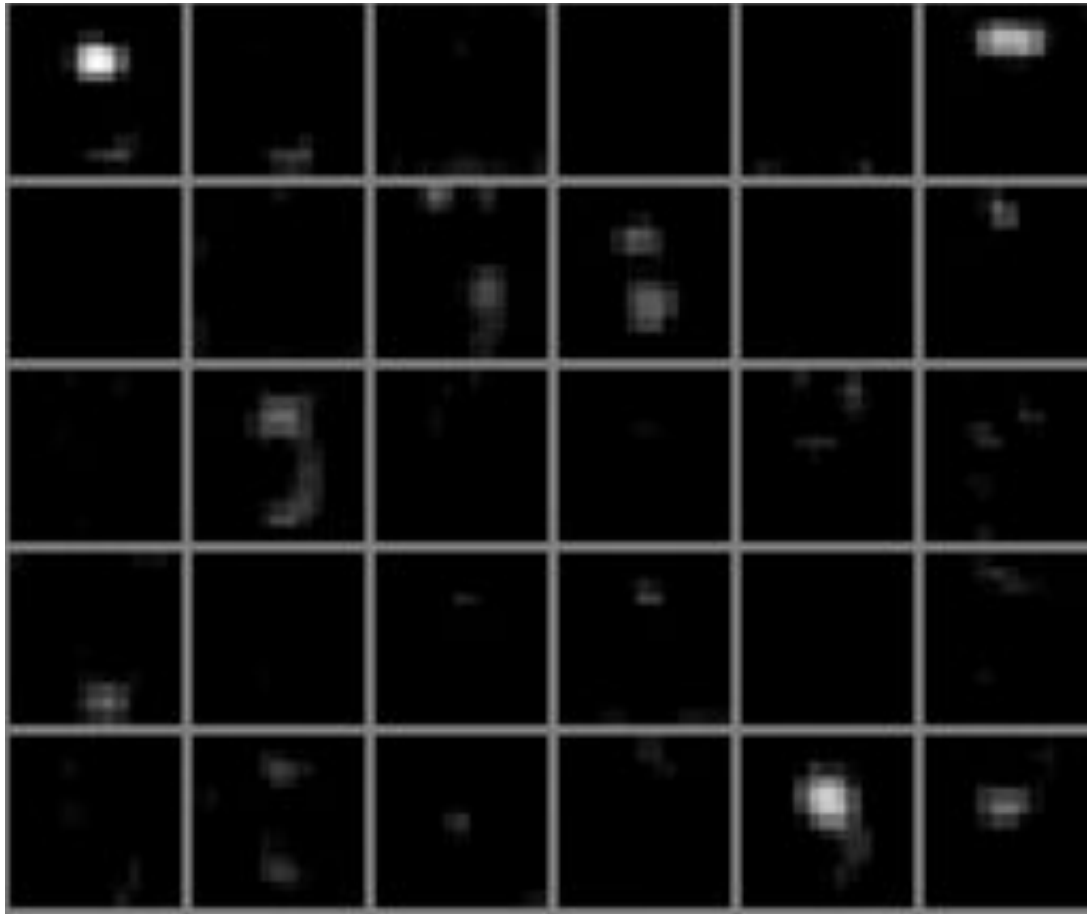
Input
image



conv1

Visualization of Feature Activations (2D)

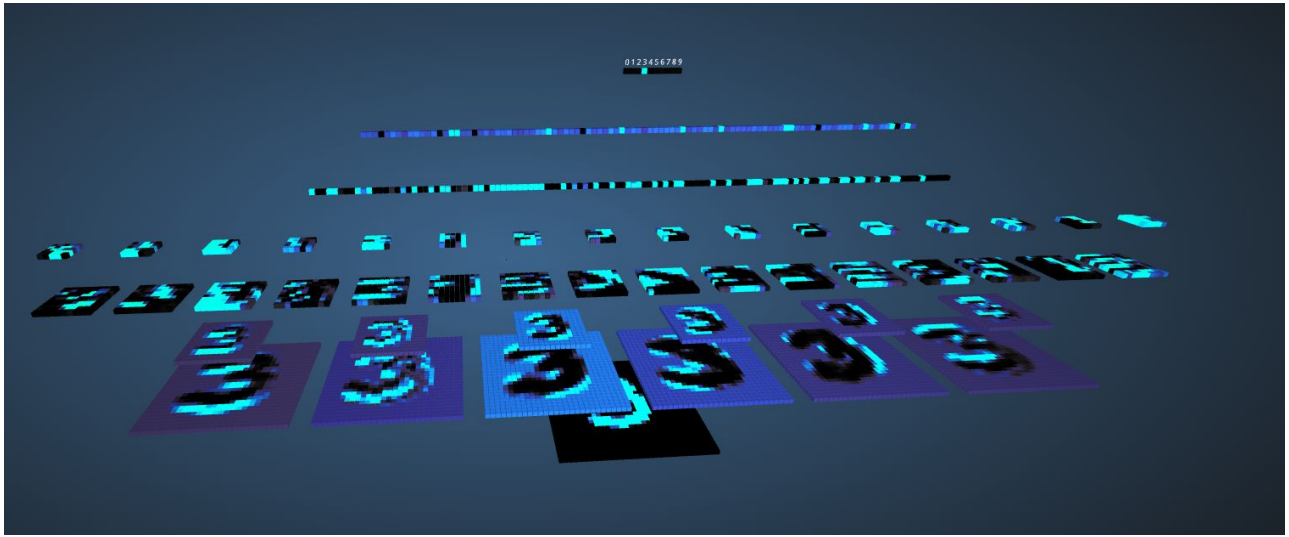
Input
image



conv5

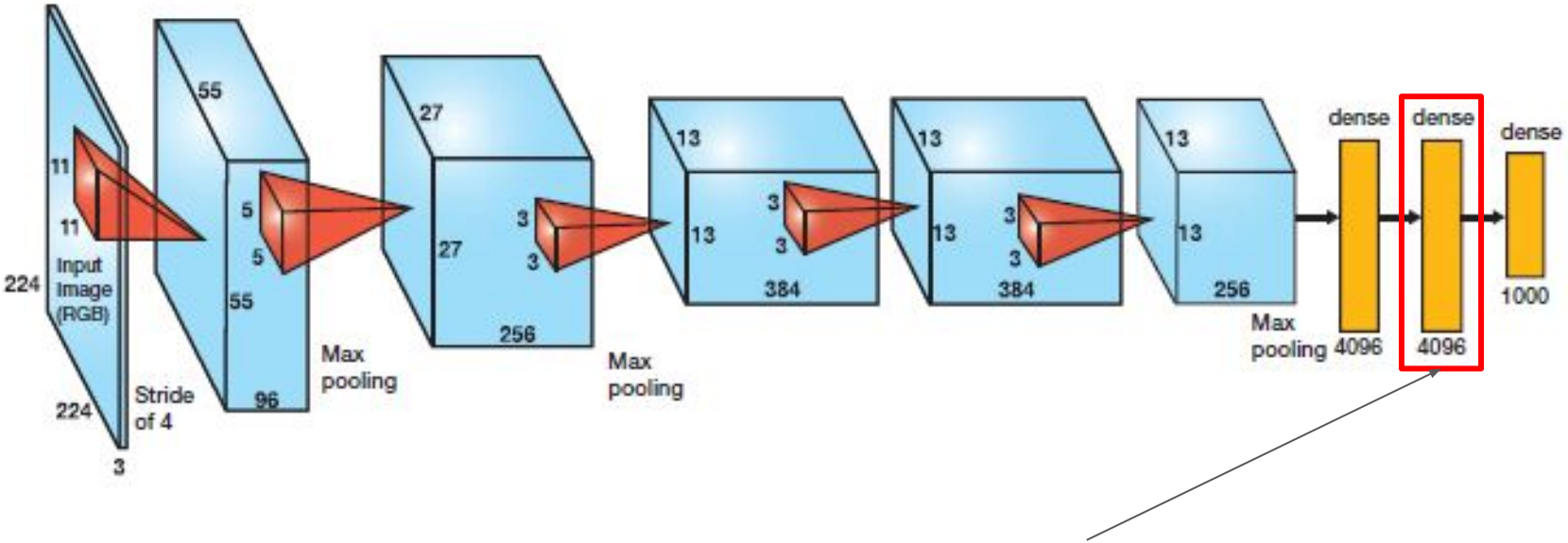
Visualization of Feature Activations (2D)

Demo: 3D Visualization of a Convolutional Neural Network



Harley, Adam W. "An Interactive Node-Link Visualization of Convolutional Neural Networks." In Advances in Visual Computing, pp. 867-877. Springer International Publishing, 2015.

Visualization of Feature Activations (any dimension)

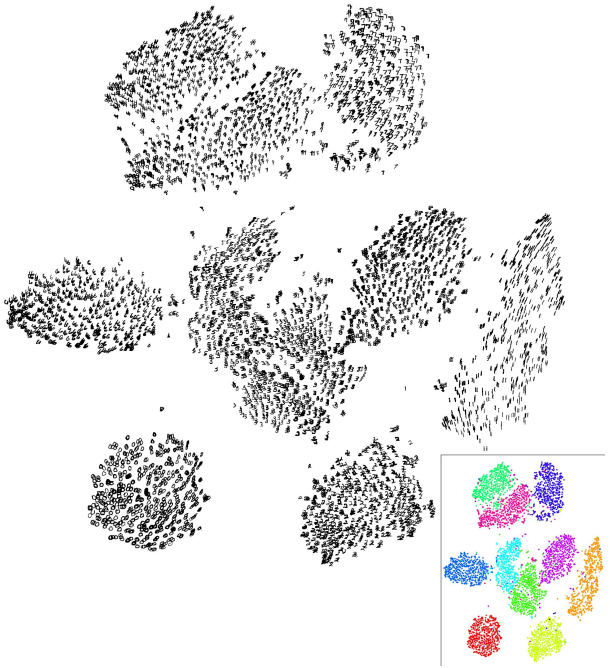


Can be used with features from layer before classification

Dimensionality reduction: t-SNE

t-SNE:
Embeds high dimensional data points (i.e. feature maps) so that pairwise distances are preserved in a local 2D neighborhoods.

Example: 10 classes from MNIST dataset



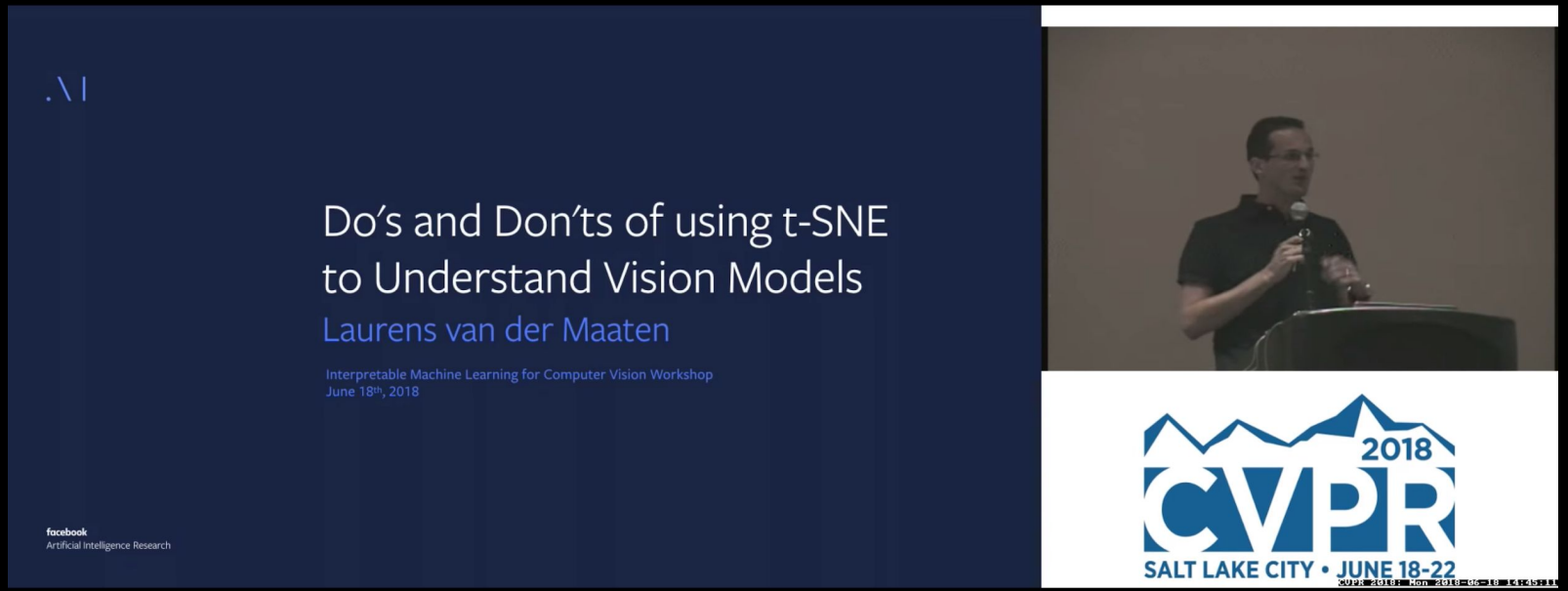
#t-sne Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. Nov (2008): 2579-2605. [Python scikit-learn]

Dimensionality reduction of Feature Activations

CNN codes:
t-SNE on fc7 features from AlexNet.

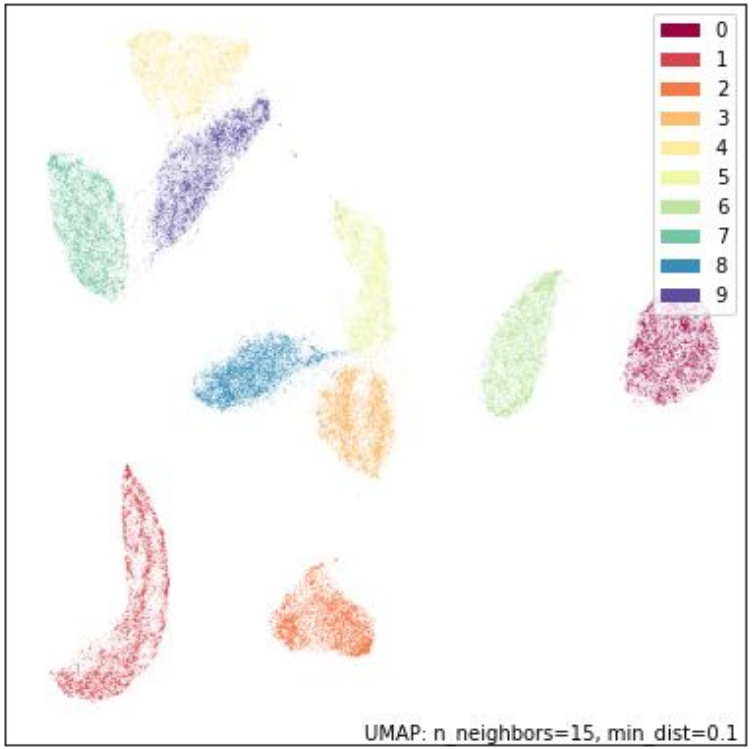
Source: Andrey Karpath (Stanford University)

#t-sne Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9, no. Nov (2008): 2579-2605. [Python scikit-learn]



Maaten, Laurens van der @ CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision.

Dimensionality reduction: UMAP



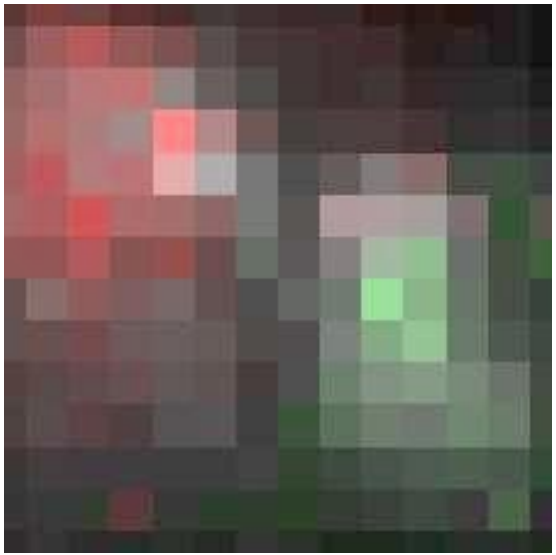
#UMAP McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29).

Interpretability: Attribution

- Visualization
- **Attribution**
  - Receptive Field of the Highest Activations
  - Activation Changes after Occlusions
  - Fully Connected into Convolutional layers
  - Class Activation Maps (CAMs)
  - Gradient-based
- Feature visualization

Attribution

Attribution studies what part of an example is responsible for the network activating a particular way.



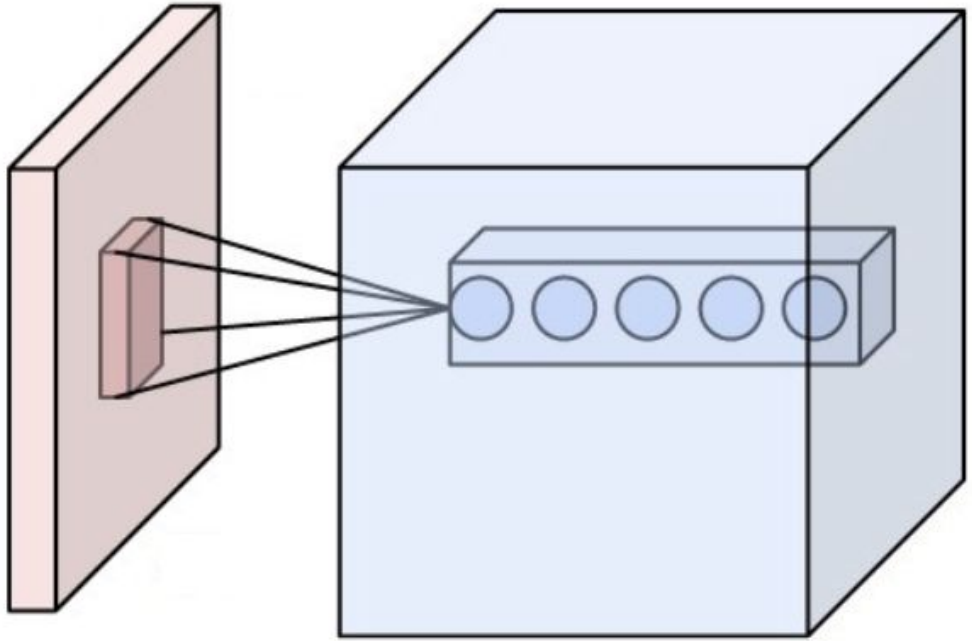
Chris Olah, Alexander Mordvintsev, Ludwig Schubert, "Feature Visualization". Distill.pub (November 2017)

Interpretability: Attribution

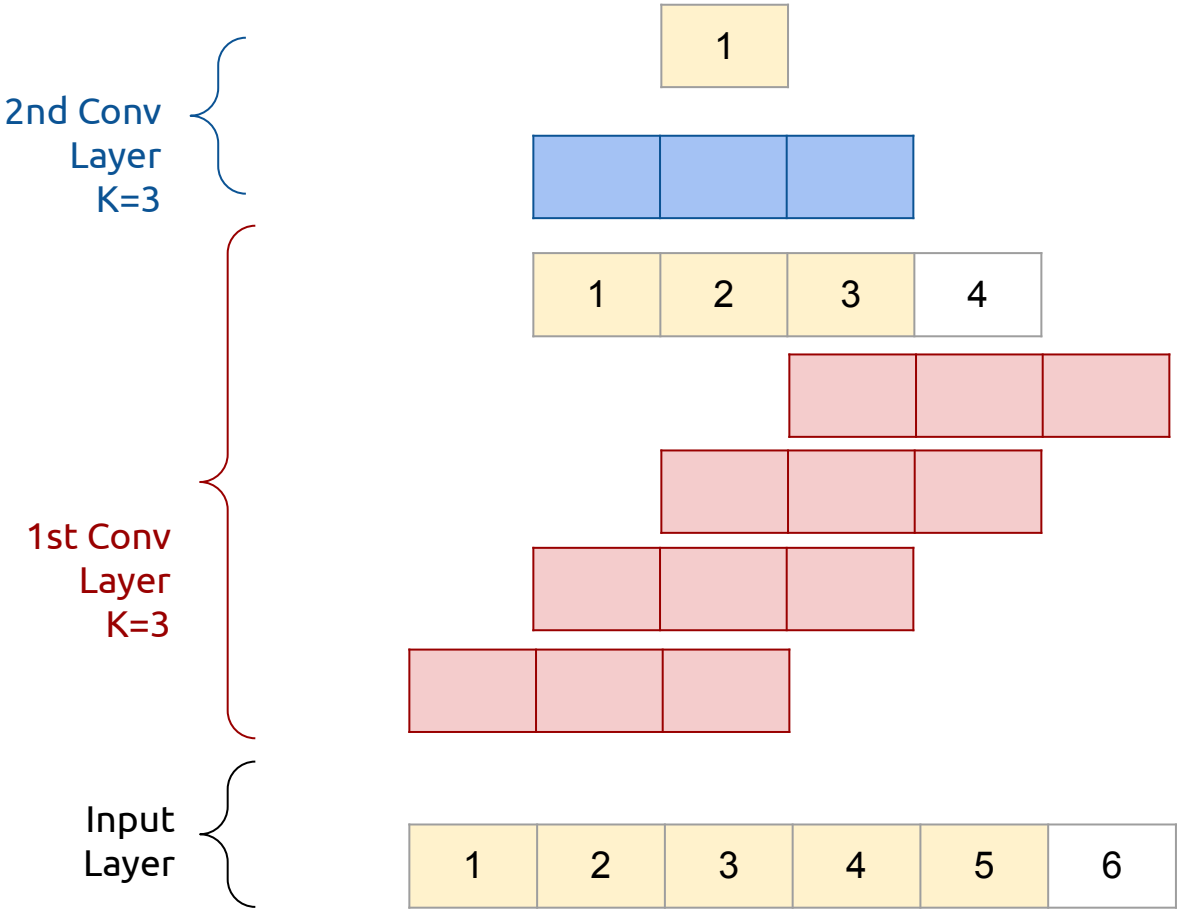
- Visualization
- **Attribution**
  - **Receptive Field of the Highest Activations**
  - Activation Changes after Occlusions
  - Fully Connected into Convolutional layers
  - Class Activation Maps (CAMs)
  - Gradient-based
- Feature visualization

Reminder: Receptive Field

Receptive field: Part of the input that is visible to a neuron. It increases as we stack more convolutional layers (i.e. neurons in deeper layers have larger receptive fields).



Receptive Field



Receptive Field of Highest Activations

Visualize the receptive field of a neuron over those images that activate this neuron the most

Girshick et al. Rich feature hierarchies for accurate object detection and semantic segmentation. CVPR 2014

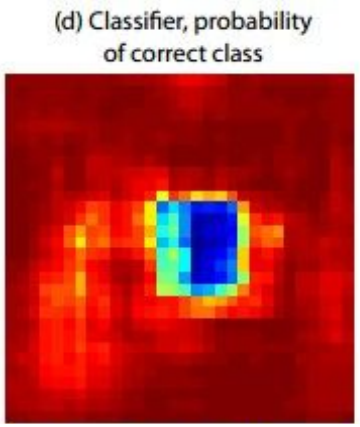
Interpretability: Attribution

- Visualization
- **Attribution**
  - Receptive Field of the Highest Activations
  - **Activation Changes after Occlusions**
  - Fully Connected into Convolutional layers
  - Class Activation Maps (CAMs)
  - Guided Backpropagation
- Feature visualization

Activation Changes after Occlusions (global scale)

1. Iteratively forward the same image through the network, occluding a different region at a time.

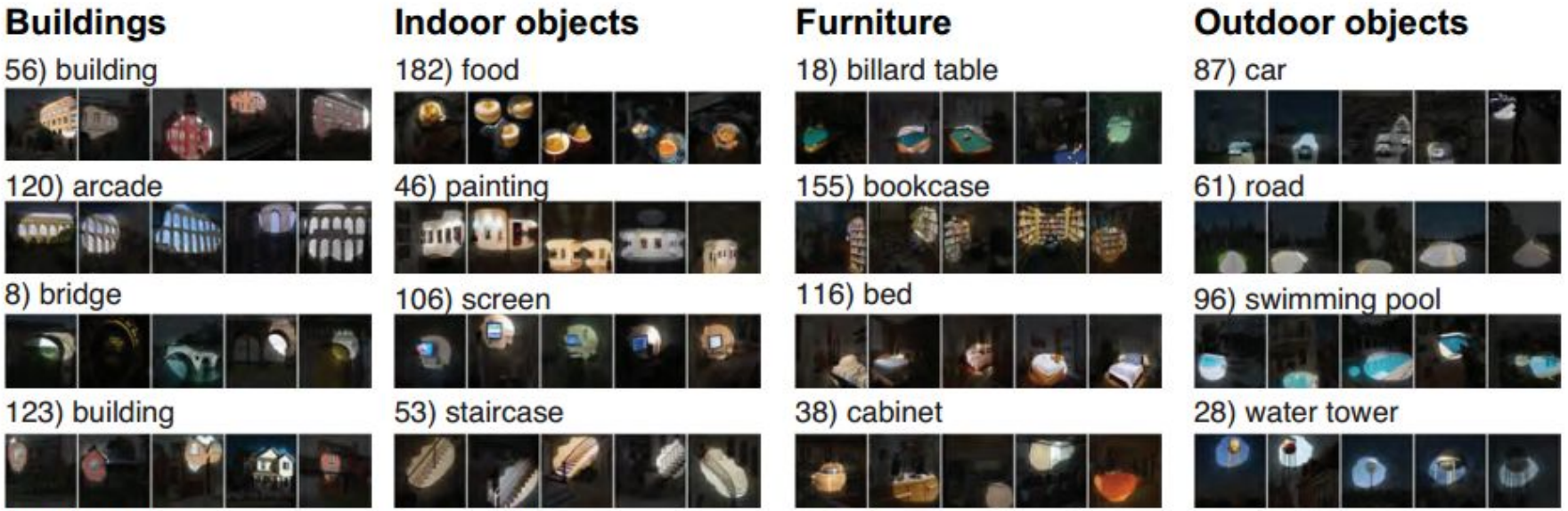
2. Keep track of the probability of the correct class with respect to the position of the occluder



Zeiler and Fergus. Visualizing and Understanding Convolutional Networks. ECCV 2015

Activation Changes after Occlusions (global scale)

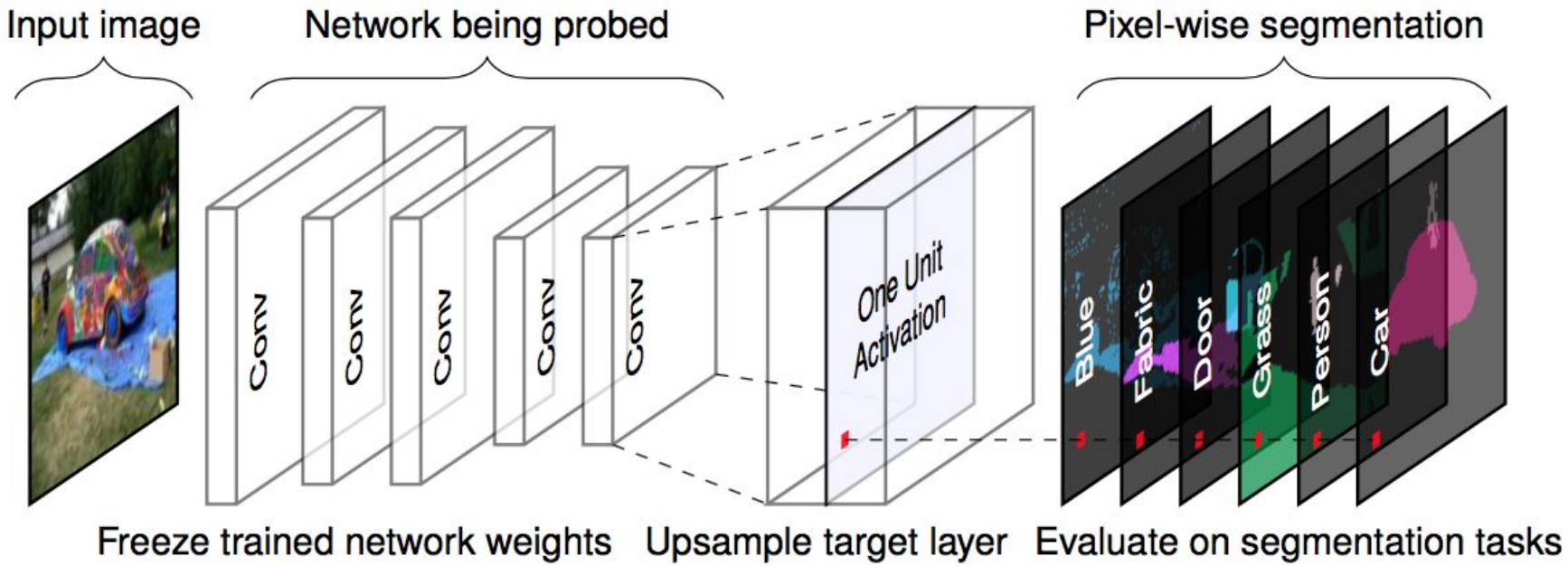
The changes in activations can be observed in any layer. This allowed identifying some neurons as weak object detectors, trained with image labels.



Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Object detectors emerge in deep scene CNNs." ICLR 2015.

Activation Changes after Occlusions (pixel scale)

Same idea, but label each neuron (unit) by means of an image dataset with pixel-level annotations.

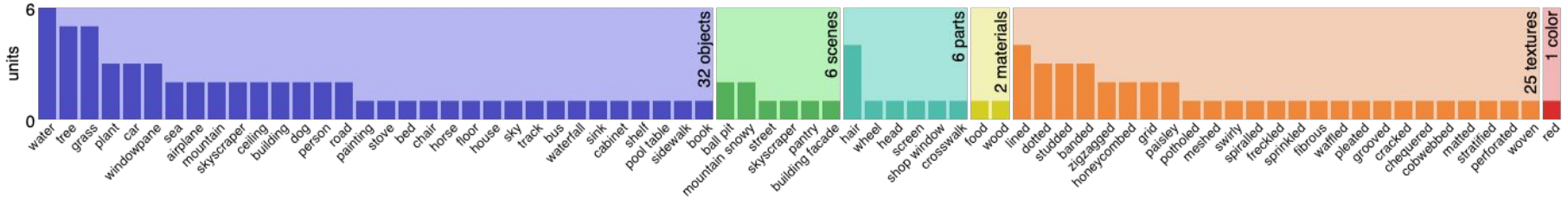


Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. CVPR 2017.

Activation Changes after Occlusions (pixel scale)

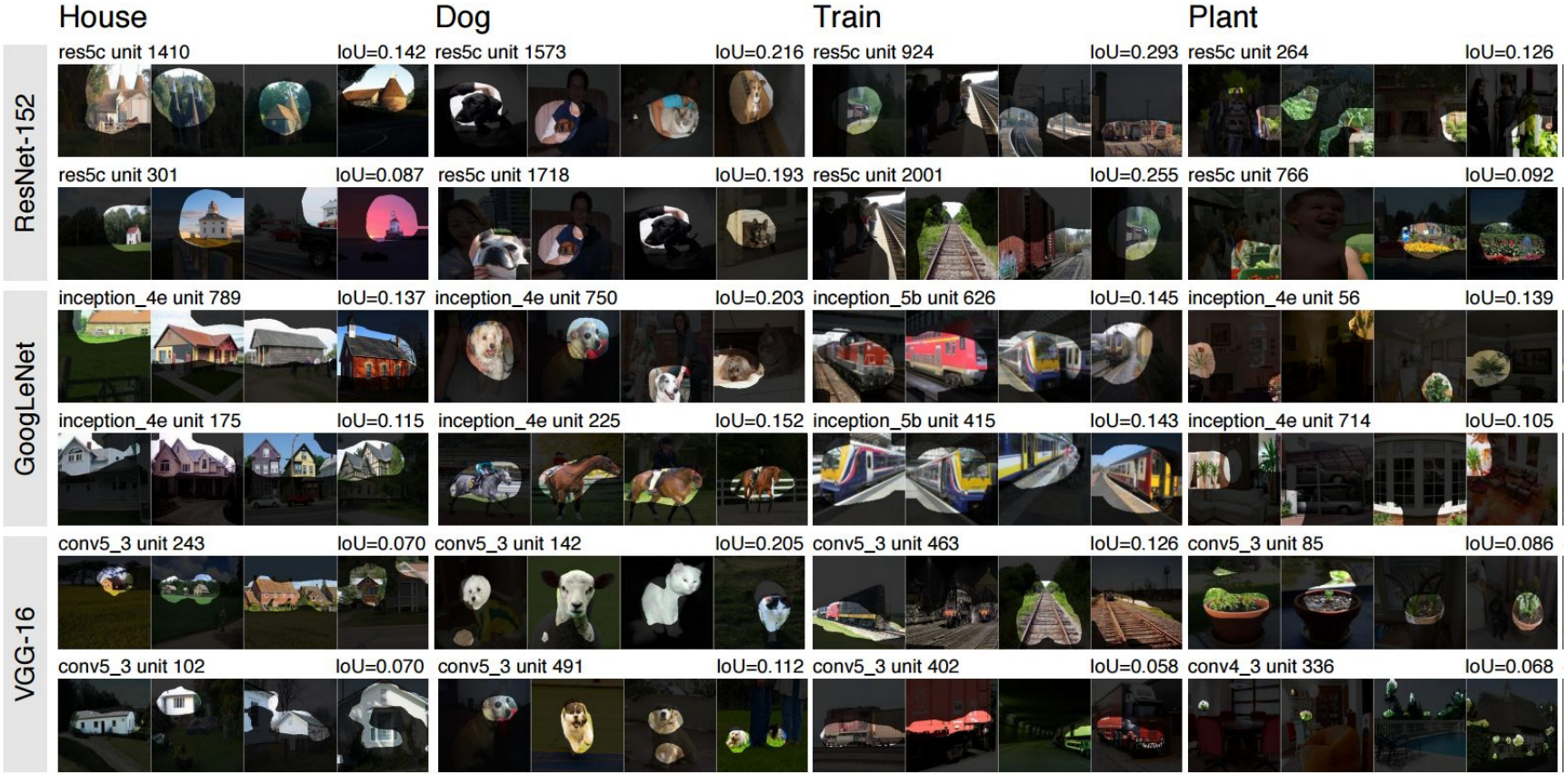
By measuring the concept that best matches each neuron (unit), Net Dissection can break down the types of concepts represented in a layer.

Here the 256 units of AlexNet conv5 trained on Places dataset represent many objects and textures, as well as some scenes, parts, materials, and a color:



Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. CVPR 2017.

Activation Changes after Occlusions (pixel scale)



Bau, D., Zhou, B., Khosla, A., Oliva, A., & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. CVPR 2017.

Interpretability: Attribution

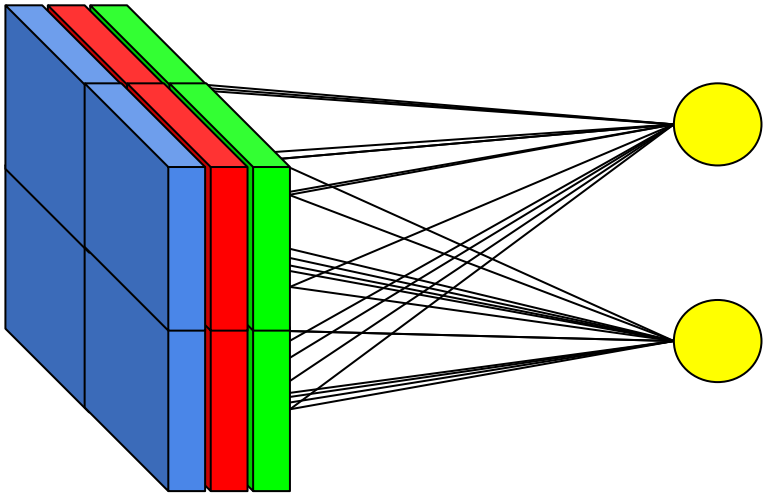
- Visualization
- **Attribution**
  - Receptive Field of the Highest Activations
  - Activation Changes after Occlusions
  - **Fully Connected into Convolutional layers**
  - Class Activation Maps (CAMs)
  - Guided Backpropagation
- Feature visualization

Fully Connected into Convolutional Layers

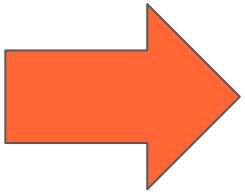
By redefining fully connected neurons into convolutional neurons...

3x2x2 tensor
(RGB image of 2x2)

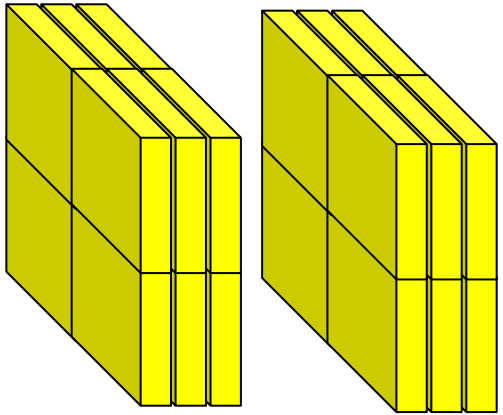
2 fully connected
neurons



3x2x2 * 2 weights



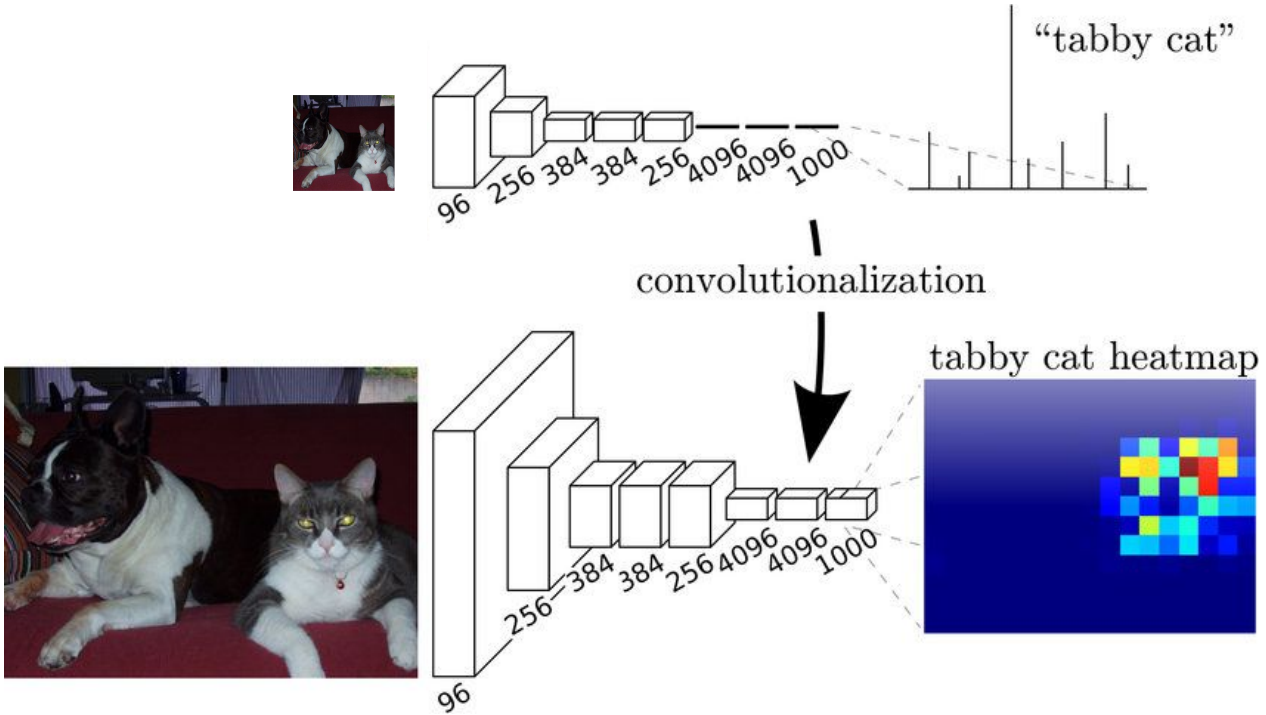
2 convolutional filters of 3 x 2 x 2
(same size as input tensor)



3x2x2 * 2 weights

Fully Connected into Convolutional Layers

...a model trained for image classification on low-definition images can provide local response when fed with high-definition images.



Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." CVPR 2015. (original figure has been modified)

Fully Connected into Convolutional Layers

The FC to Conv redefinition allows generating heatmaps of the class prediction over the input images.

Interpretability: Attribution

- Visualization
- **Attribution**
  - Receptive Field of the Highest Activations
  - Activation Changes after Occlusions
  - Fully Connected into Convolutional layers
  - **Class Activation Maps (CAMs)**
  - Guided Backpropagation
- Feature visualization

Reminder: Global Average Pooling (GAP)

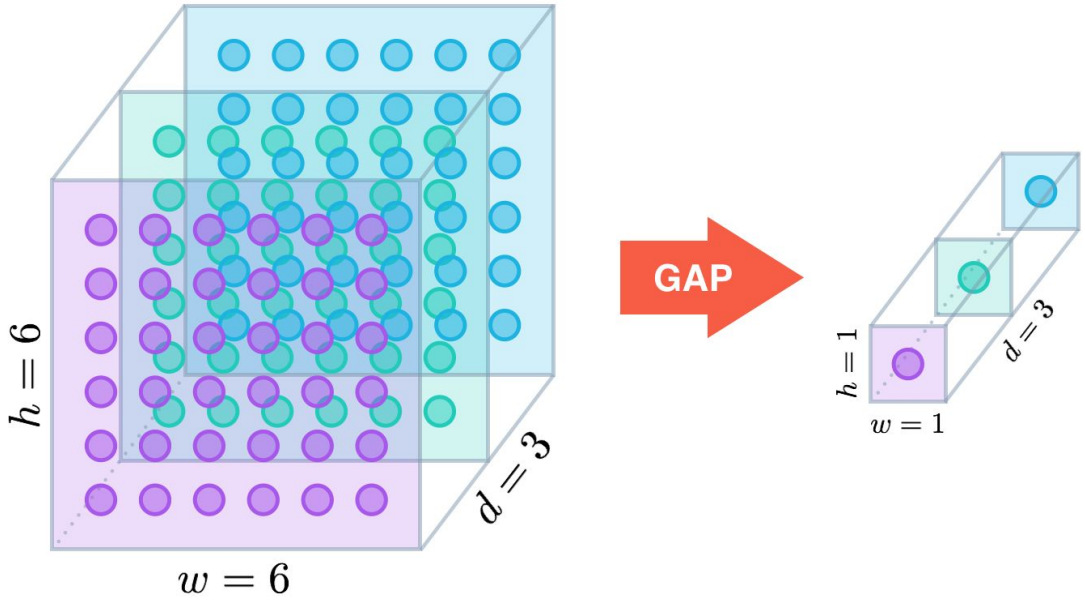
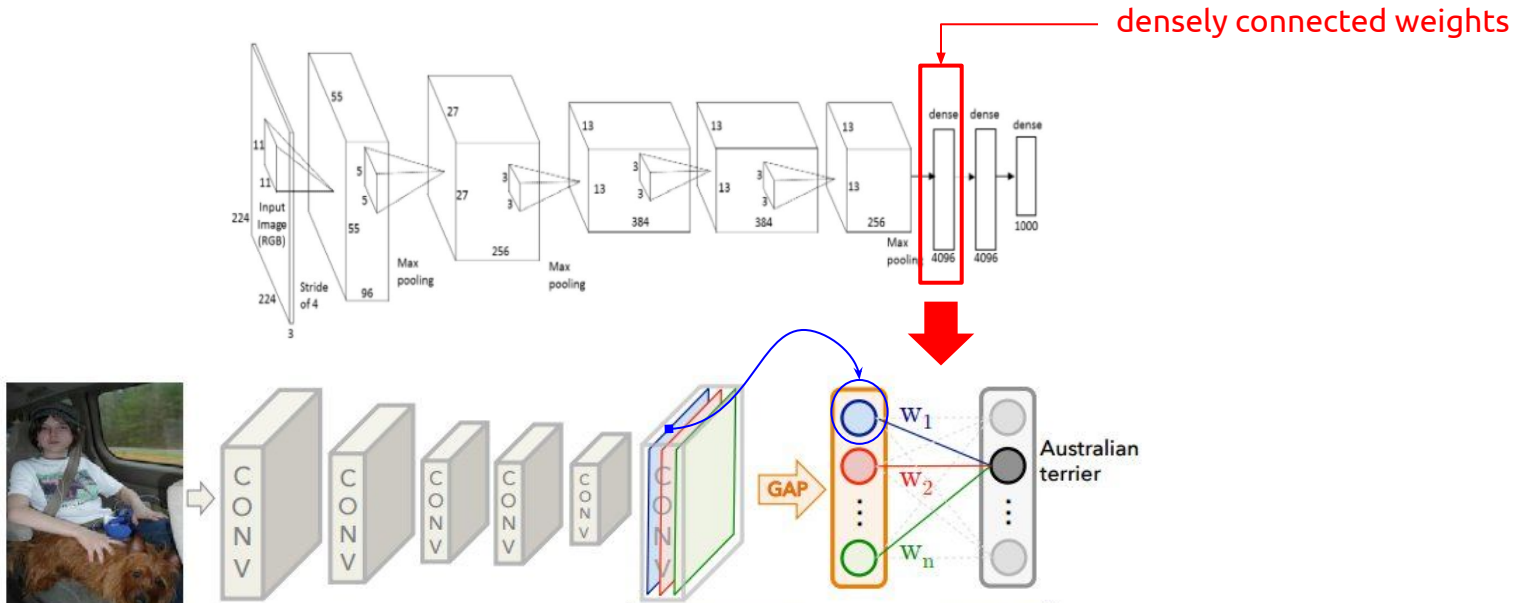


Figure:
Alexis Cook

#NiN Lin, Min, Qiang Chen, and Shuicheng Yan. "Network in network." *ICLR 2014.*

Class Activation Maps (CAMs)

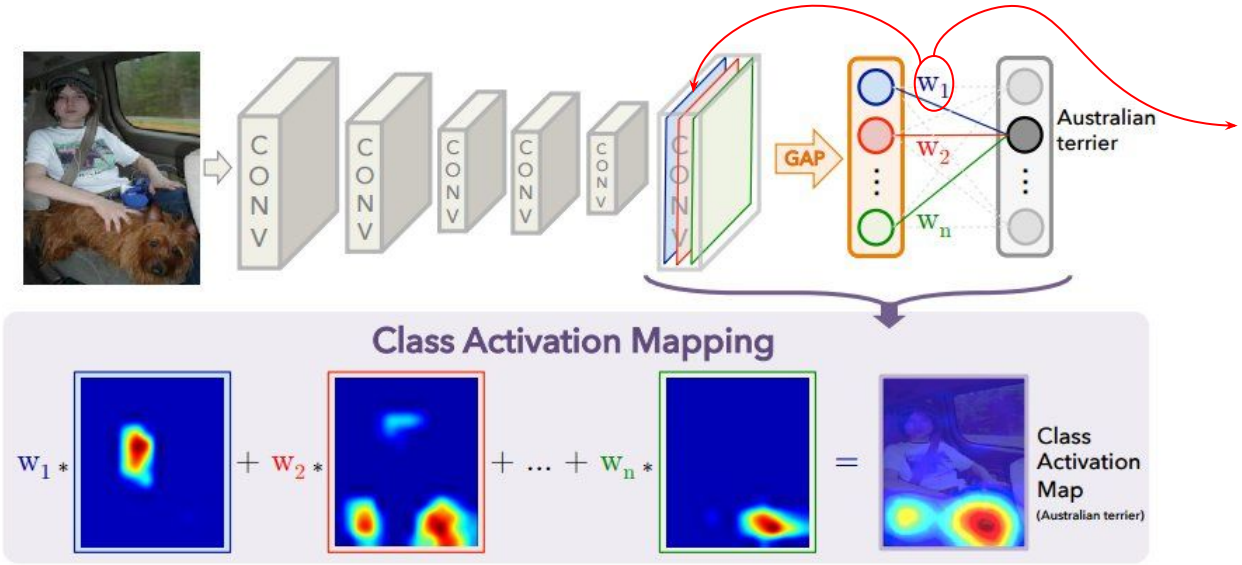
Change in a classic CNN+MLP architecture: Replace FC layer after last conv with Global Average Pooling (GAP), which corresponds to averaging per channel.



#CAM Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." CVPR 2016

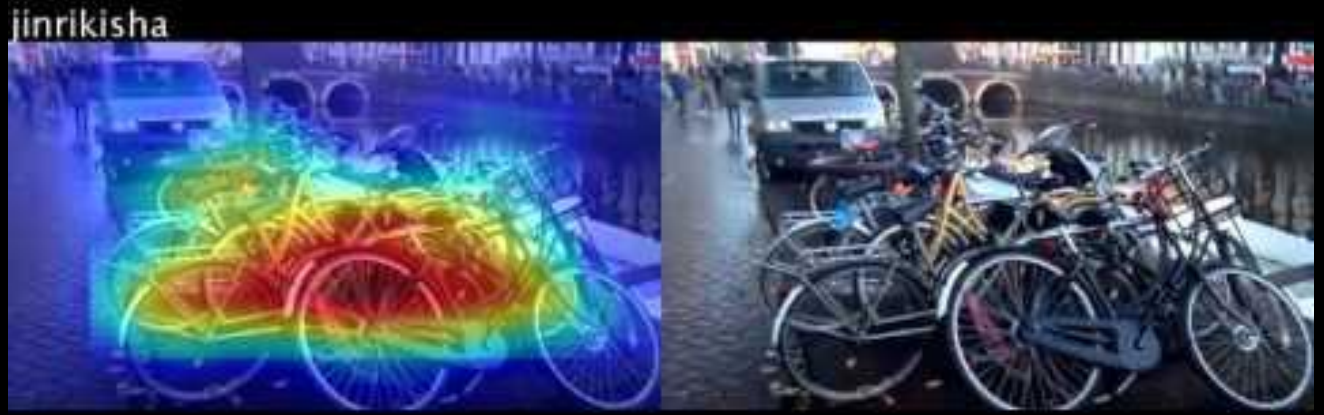
Class Activation Maps (CAMs)

Weighted Fusion of Feature Maps : The neuron weights define the contribution of each channel in the previous layer



contribution of the *blue* channel in the conv layer to the prediction of “australian terrier”

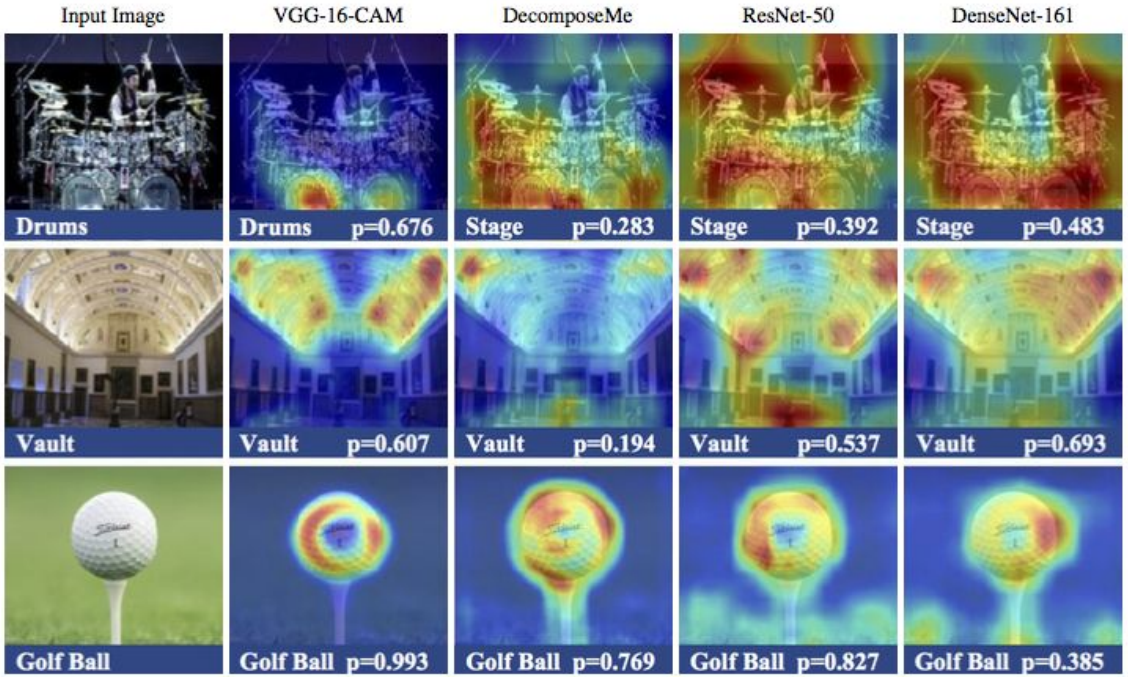
#CAM Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." CVPR 2016



#CAM Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." CVPR 2016

Class Activation Maps (CAMs)

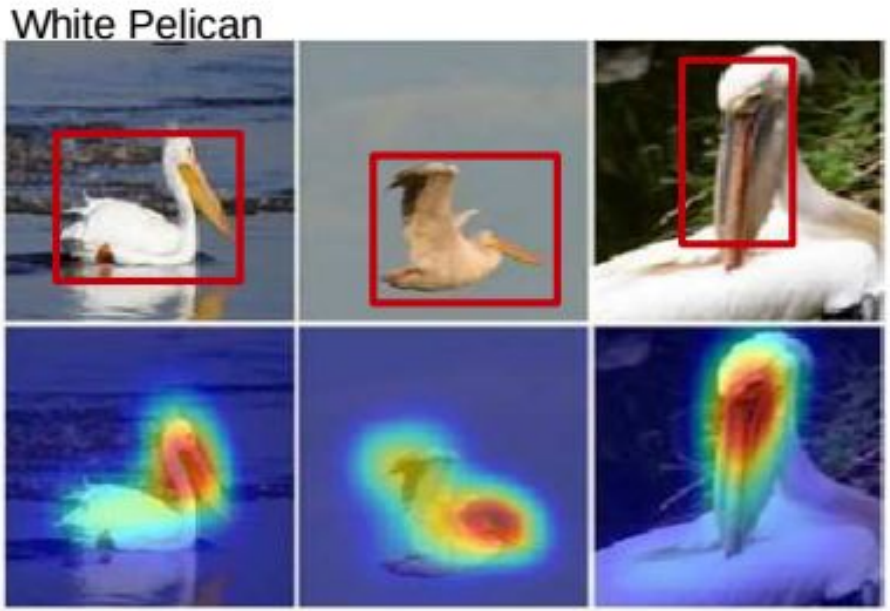
Different architectures generate different CAMs.

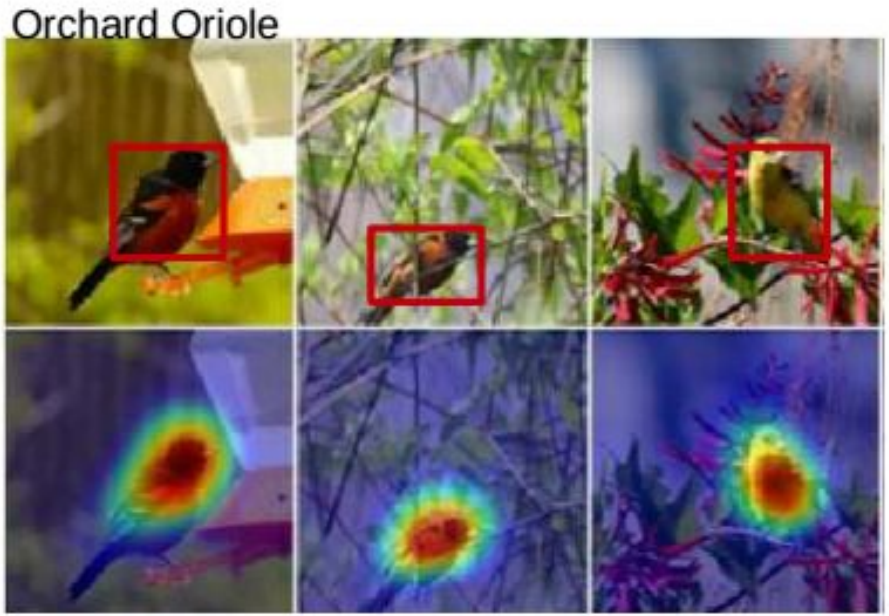


Jimenez, Albert, Jose M. Alvarez, and Xavier Giro-i-Nieto. "Class-weighted convolutional features for visual instance search." BMVC 2017.

Class Activation Maps (CAMs)

Applications of CAMs: Rough object localization with weak labels (image).





Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." CVPR 2016

Class Activation Maps (CAMs)

Applications of CAMs: Spatial feature weighting for object retrieval.

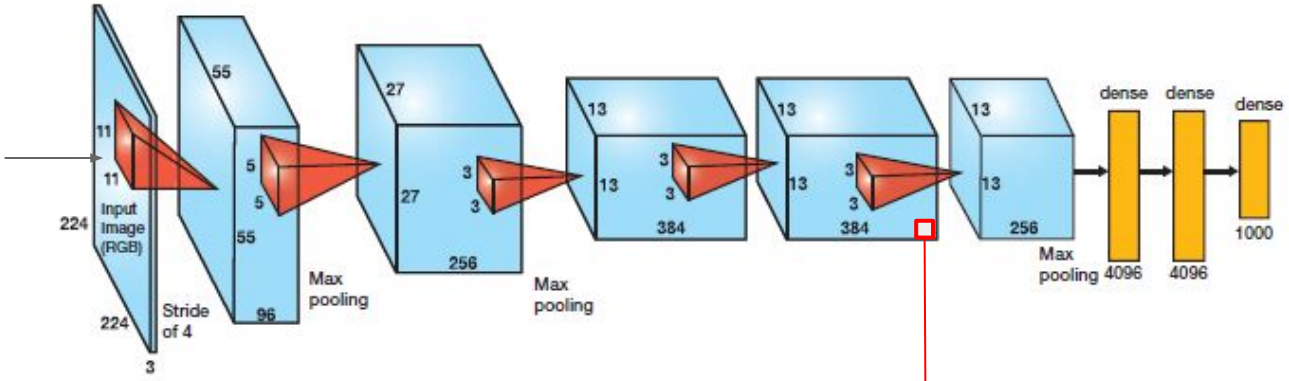
Jimenez, Albert, Jose M. Alvarez, and Xavier Giro-i-Nieto. "Class-weighted convolutional features for visual instance search." BMVC 2017.

Interpretability: Attribution

- Visualization
- **Attribution**
  - Receptive Field of the Highest Activations
  - Activation Changes after Occlusions
  - Fully Connected into Convolutional layers
  - Class Activation Maps (CAMs)
  - **Guided Backpropagation**
- Feature visualization

Guided Backpropagation

Goal: Visualize the part of an image that mostly activates one of the neurons.

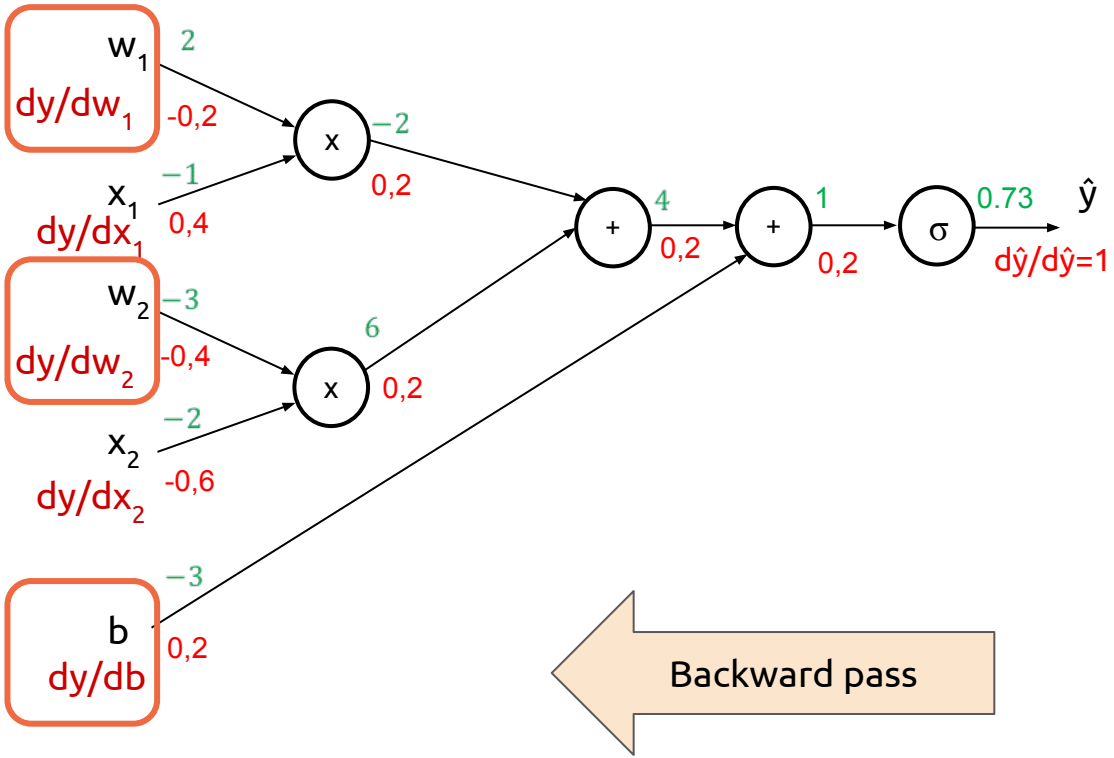


1. Forward image up to the layer that want to be interpreted (e.g. conv5)
2. Set all gradients to 0
3. Set gradient for the specific neuron we are interested in to 1
4. Backpropagate and update the values on the image (not the network parameters).

Guided Backpropagation

During NN training:

Normally, we will be interested only on the weights (w_i) and biases (b), not the inputs (x_i). The weights are the parameters to learn in our models.

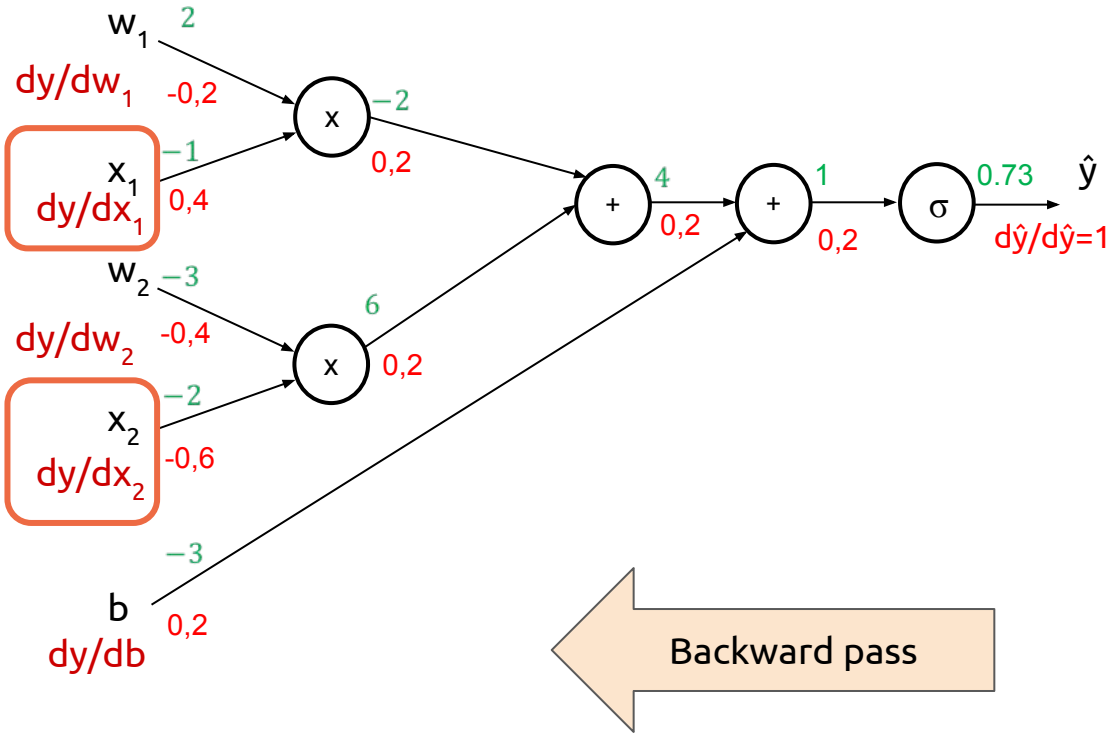


Example extracted from Andrej Karpathy's notes for CS231n from Stanford University.

Guided Backpropagation

<u>In this approach</u>:

We will be interested on the inputs (x_i), not in updating the weights (w_i) and biases (b), which are what we actually want to be able to interpret.



Example extracted from Andrej Karpathy's notes for CS231n from Stanford University.

Guided Backpropagation

Goal: Visualize the part of an image that mostly activates one of the neurons.

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M.. Striving for Simplicity: The All Convolutional Net. ICLR 2015

Guided Backpropagation

guided backpropagation

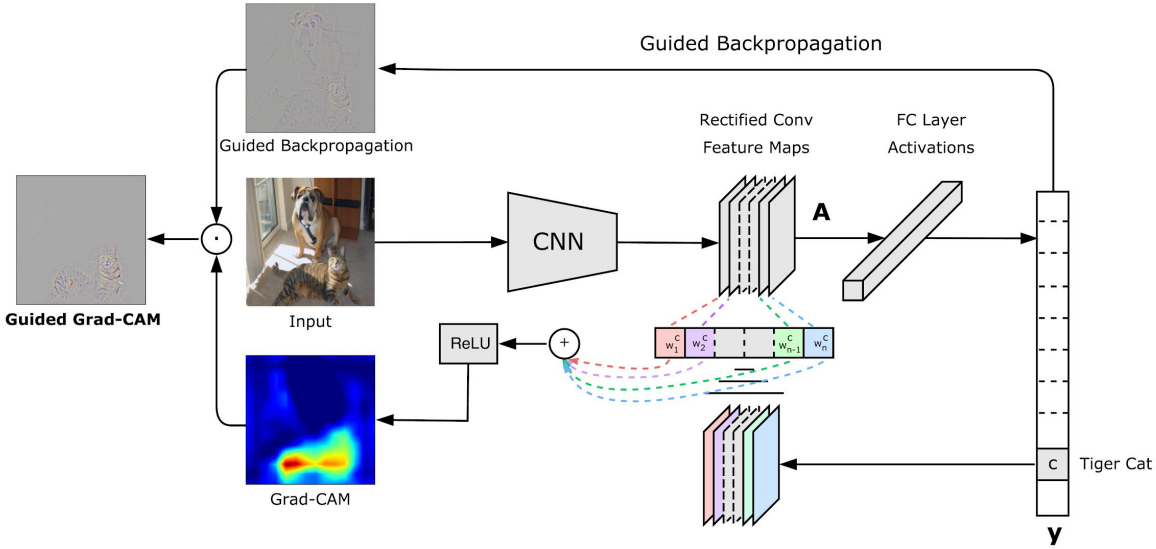
corresponding image crops

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M.. Striving for Simplicity: The All Convolutional Net. ICLR 2015

Grad-CAM

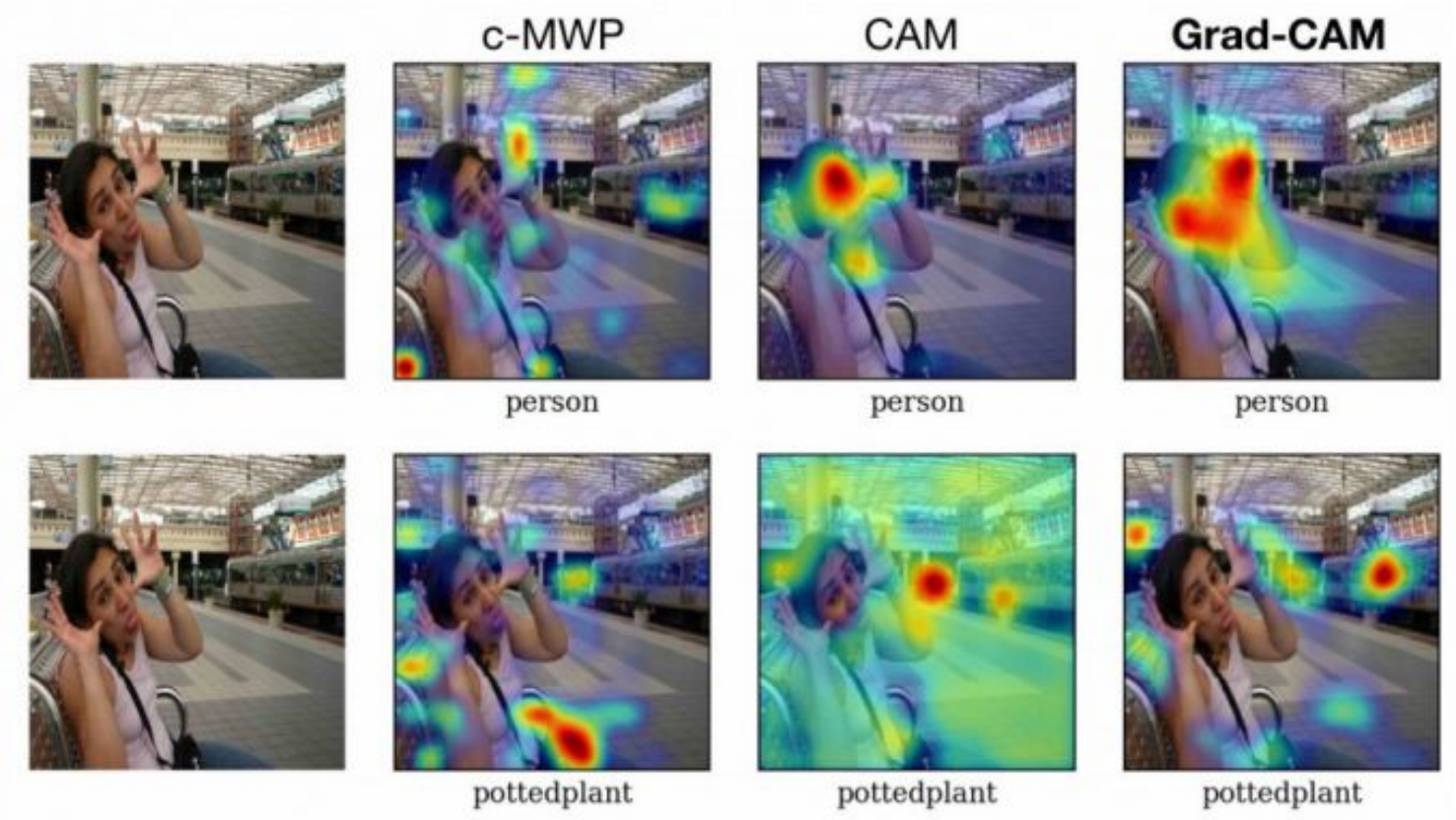
Same idea as Class Activation Maps (CAMs), but:

- No modifications required to the network
- Weight feature map by the gradient of the class wrt each channel

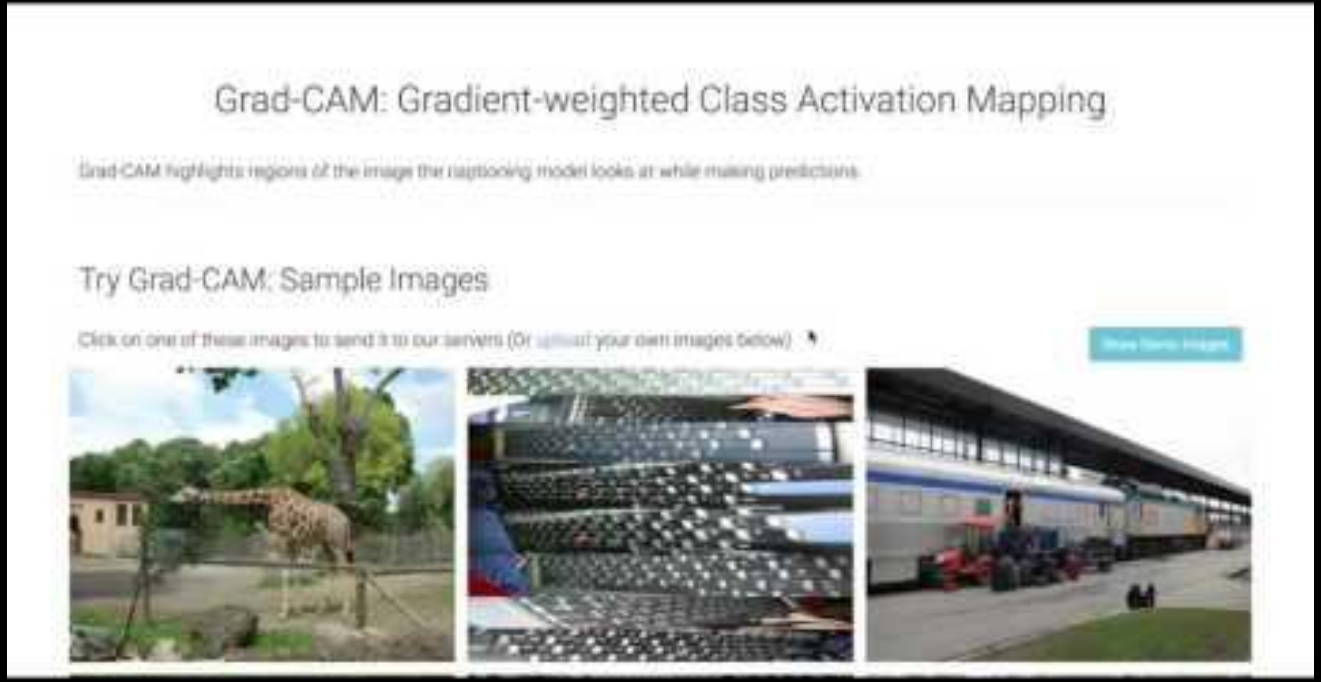


#Grad-CAM Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. ICCV 2017 [video]

Grad-CAM

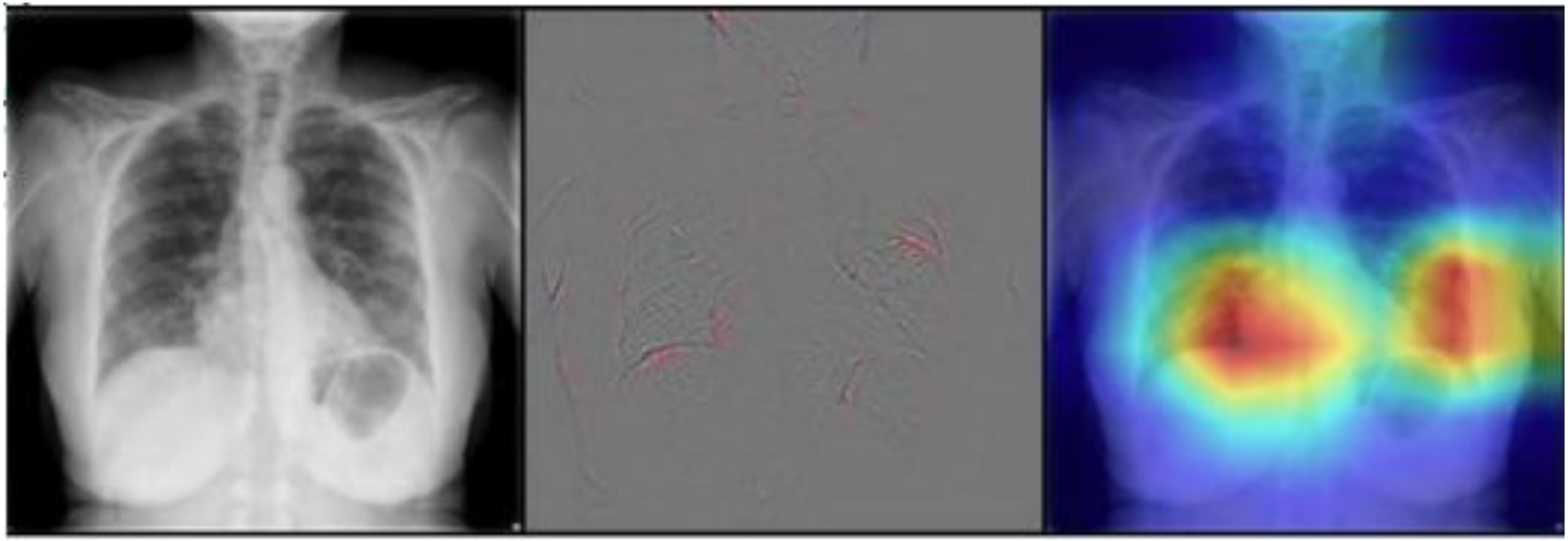


#Grad-CAM Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. ICCV 2017 [video]



#Grad-CAM Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization. ICCV 2017 [video]

Grad-CAM



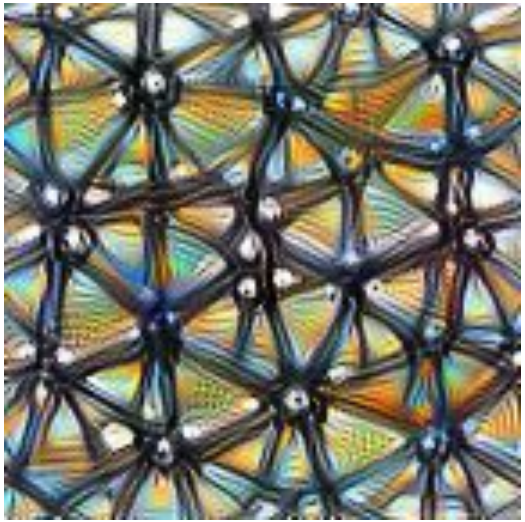
Belén del Río, Marc Boher, Borja Velasco, Toni Serena. Advised by Santi Puch., "FastConvNet". UPC School 2020

Interpretability: Attribution

- Visualization
- Attribution
- **Feature visualization**

Feature Visualization

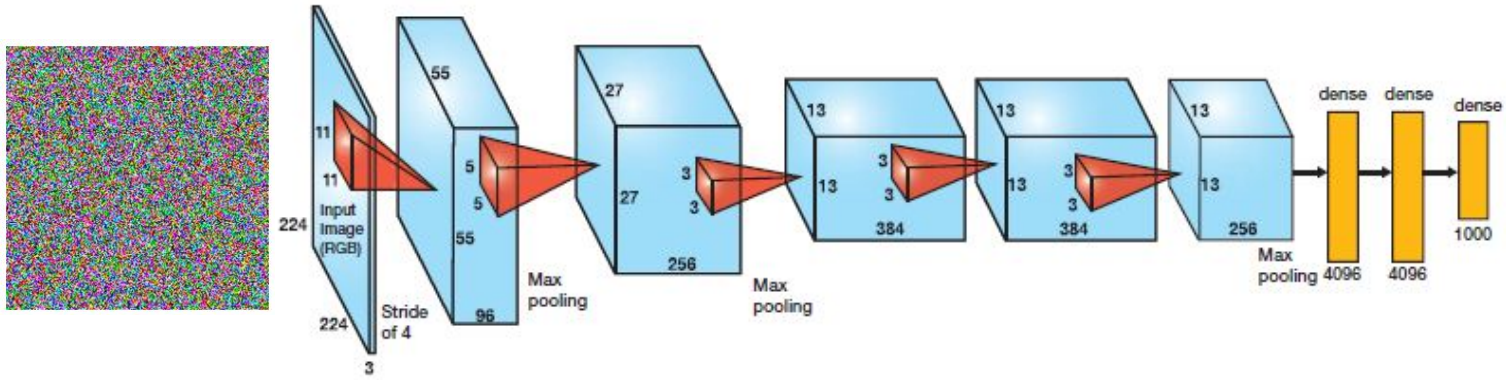
Feature visualization answers questions about what a network — or parts of a network — are looking for by generating examples.



Chris Olah, Alexander Mordvintsev, Ludwig Schubert, "Feature Visualization". Distill.pub (November 2017)

Feature Visualization

Goal: Generate an image that maximizes the activation of a neuron.

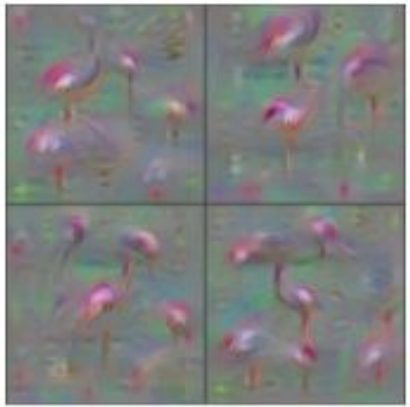


1. Forward random image
2. Set the gradient of the neuron to be [0,0,0…,1,...,0,0]
3. Backprop to get gradient on the image
4. Update image (small step in the gradient direction)
5. Repeat

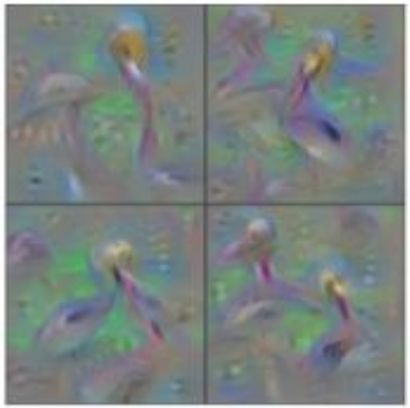
K. Simonyan, A. Vedaldi, A. Ziseramnn, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, ICLR 2014

Feature Visualization

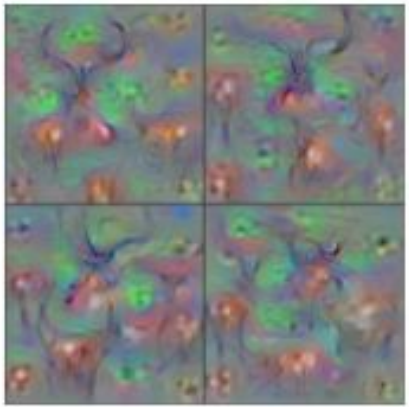
Example: Specific case of a neuron from the last layer, which is associated to a semantic class.



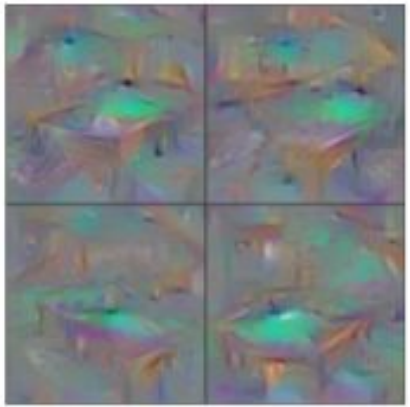
Flamingo



Pelican



Hartebeest



Billiard Table

K. Simonyan, A. Vedaldi, A. Ziseramnn, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, ICLR 2014

Feature Visualization

Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a)

Chris Olah, Alexander Mordvintsev, Ludwig Schubert, "Feature Visualization". Distill.pub (2017) [Google Colab]

Feature Visualization

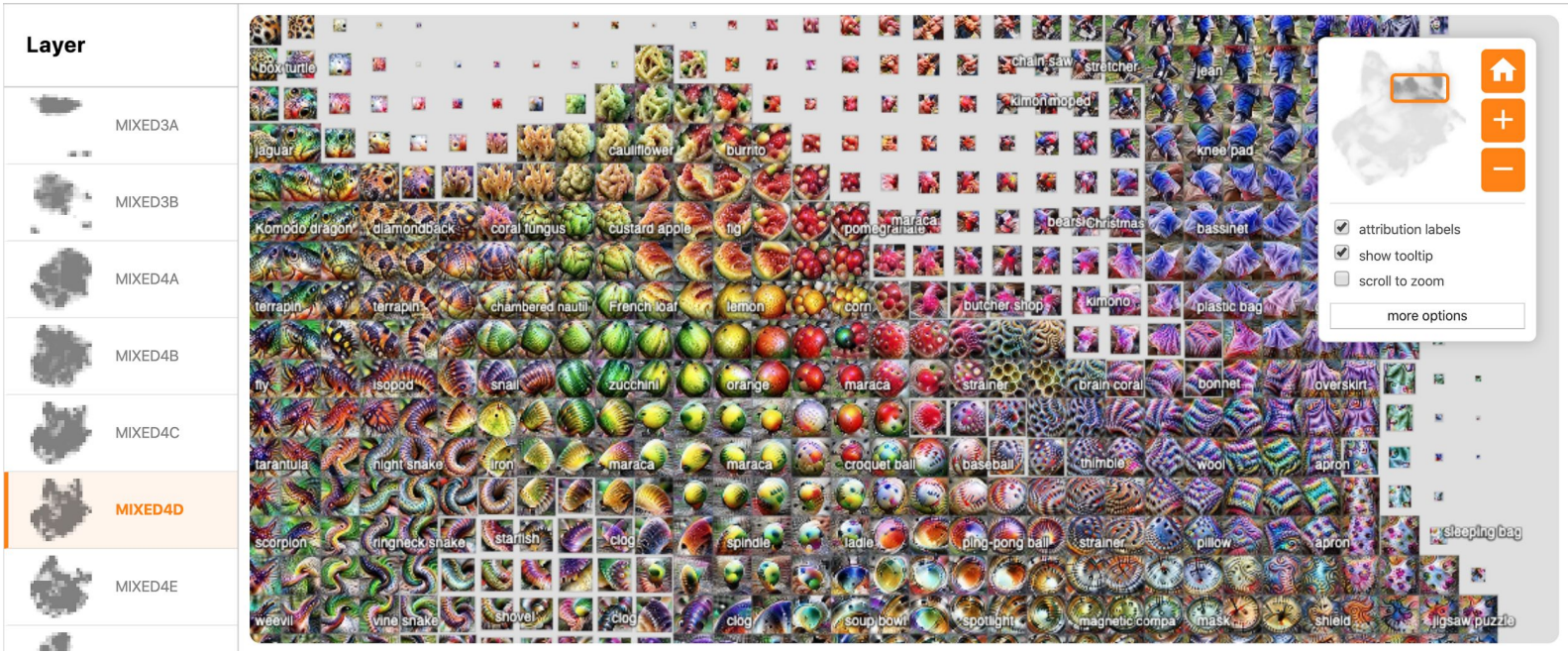
Parts (layers mixed4b & mixed4c)

Objects (layers mixed4d & mixed4e)

Chris Olah, Alexander Mordvintsev, Ludwig Schubert, "Feature Visualization". Distill.pub (2017) [Google Colab]

Activation Atlases

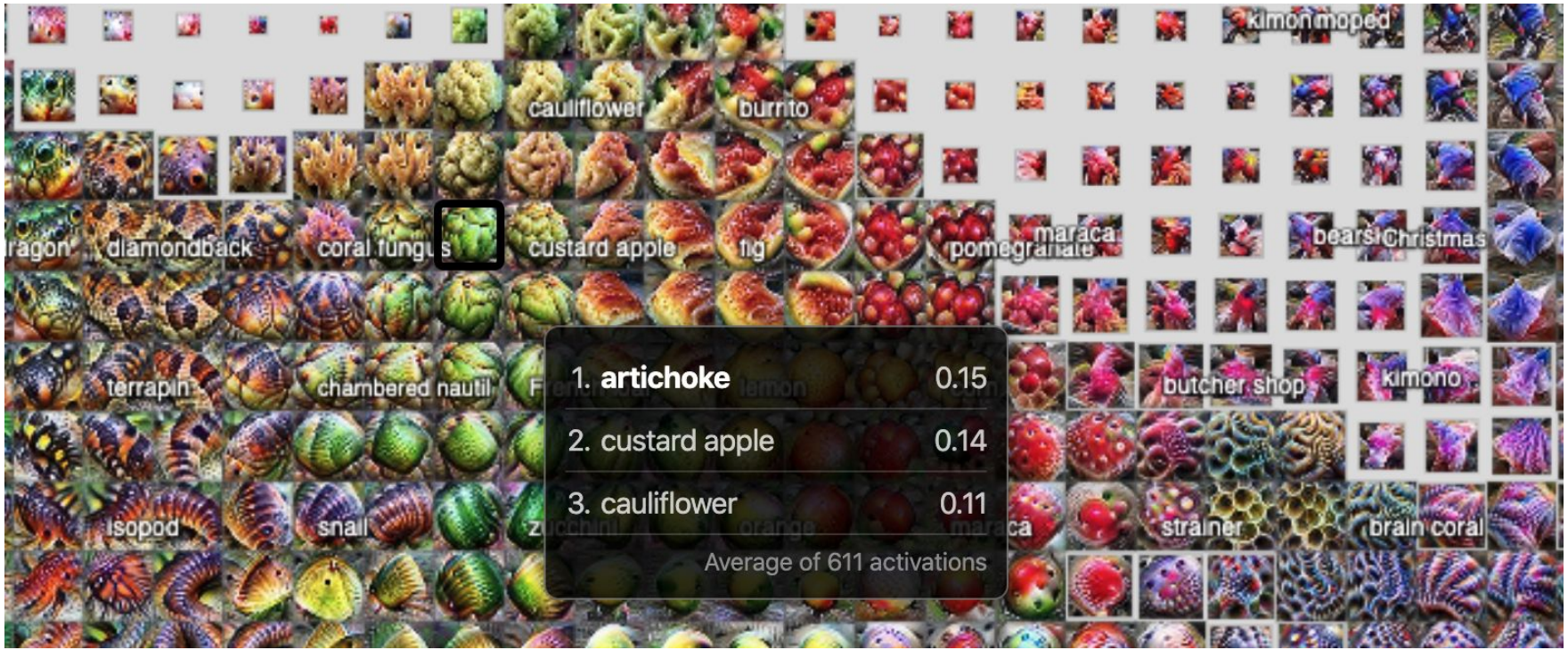
Similar to t-SNE CNN codes, but instead of showing input data, activation atlases show feature visualizations of averaged activations located in each grid cell.



Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, Chris Olah, "Exploring Neural Networks with Activation Atlases". Distill.pub (March 2019)

Activation Atlases

For each grid cell, the ImageNet classes of the contributing activations are ranked:



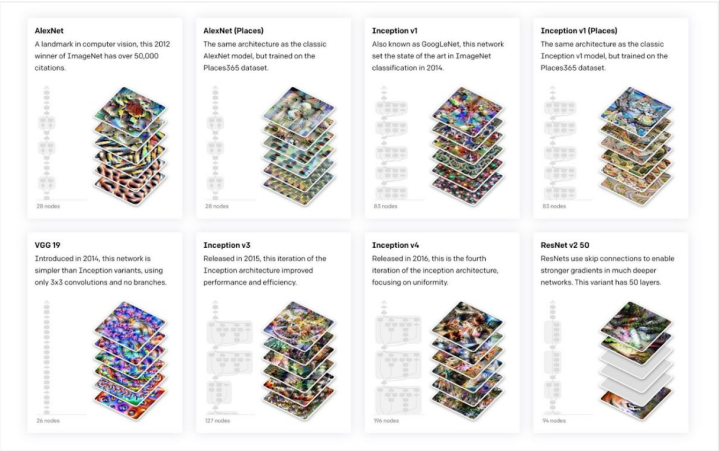
Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, Chris Olah, "Exploring Neural Networks with Activation Atlases". Distill.pub (March 2019)

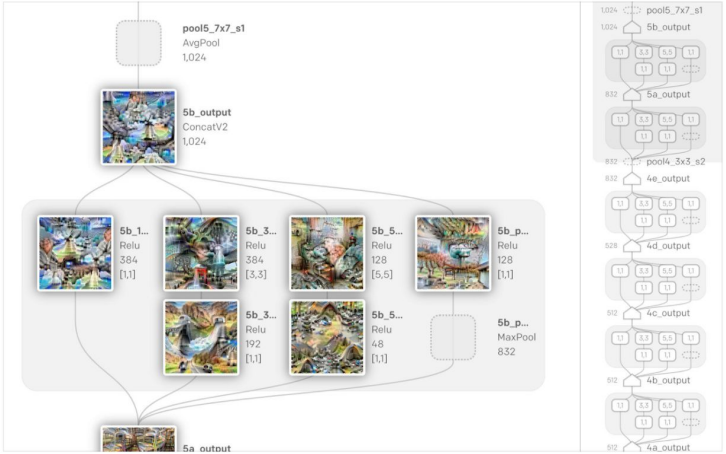
Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, Chris Olah, “Exploring Neural Networks with Activation Atlases”. Distill.pub (March 2019)

Activation Atlas + Links = OpenAI Microscope

For each grid cell, the ImageNet classes of the contributing activations are ranked:

> *InceptionV1 4c:447 is a car detector which is built from a wheel detector (4b:373) and a window detector (4b:237).*

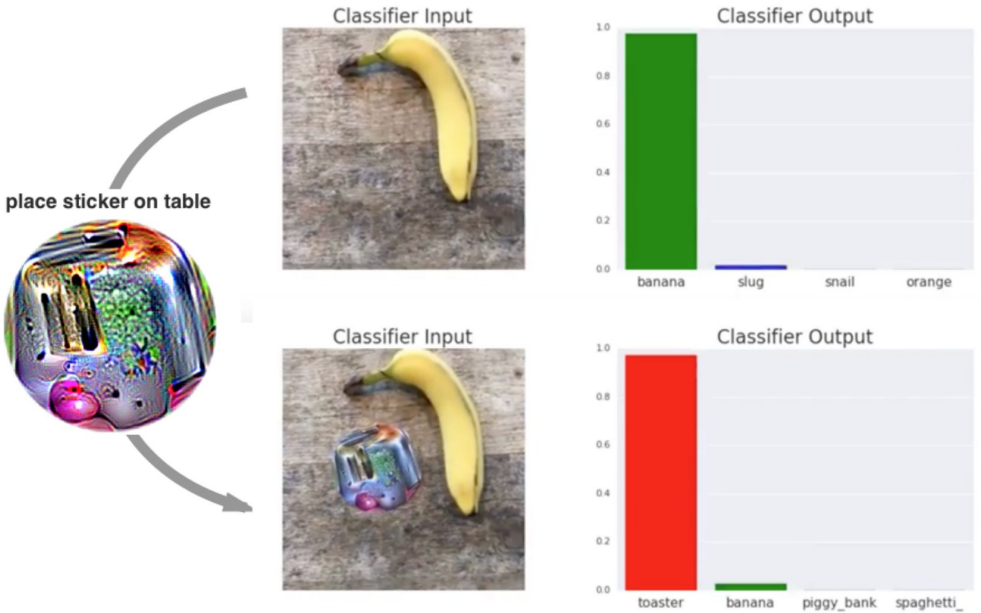




Ludwig Schubert, Michael PetrovShan Carter, Nick Cammarata, Gabriel Goh, Chris Olah, OpenAI Microscope 2020

Bonus: Adversarial Attacks (eg. patch on sticker)

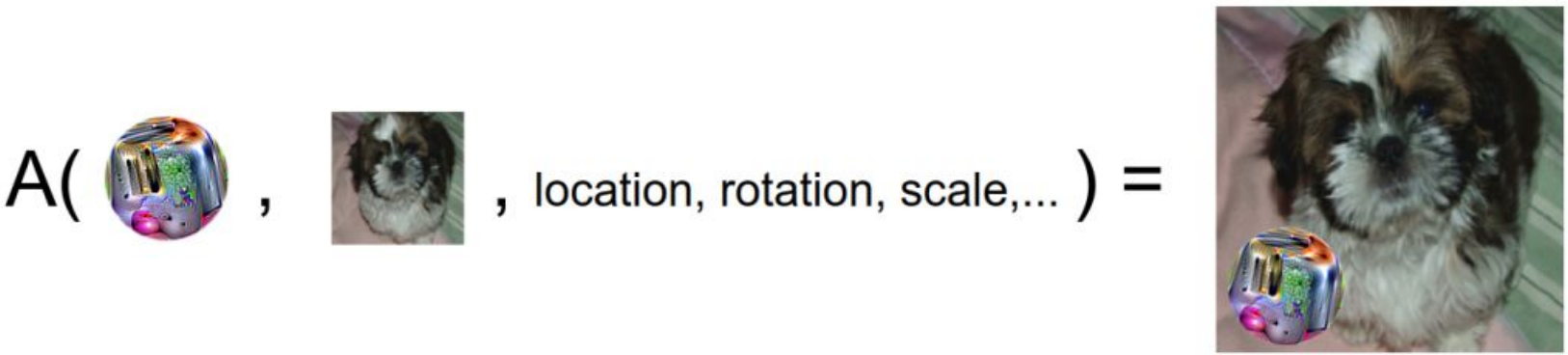
Goal: Generate a patch P that will make a Neural Network always predict class the same class for any image containing the patch.



Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. "Adversarial patch." arXiv preprint arXiv:1712.09665 (2017). [Google Colab]

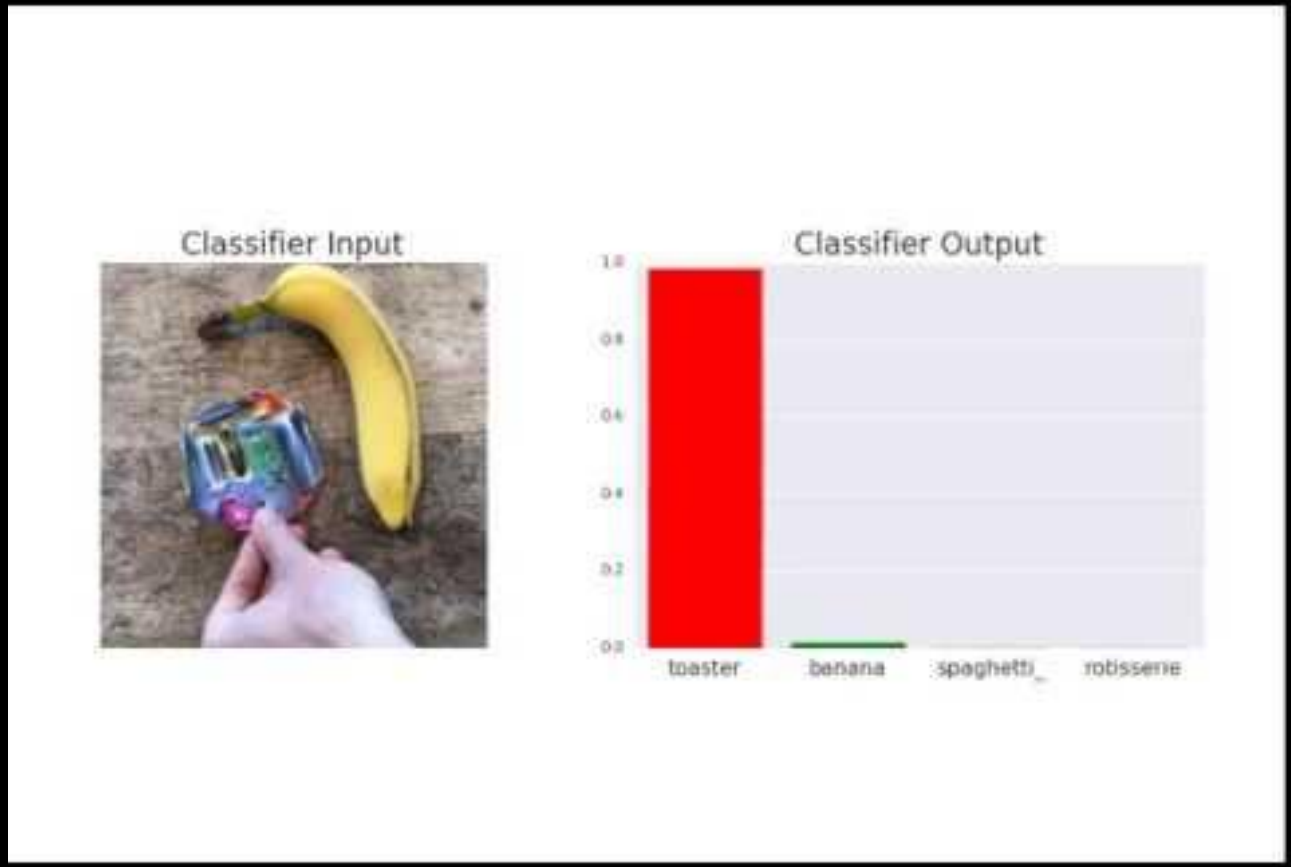
Bonus: Adversarial Attacks (eg. patch on sticker)

How: Backpropagate the gradient from the desired class to images where the patch has been inserted.



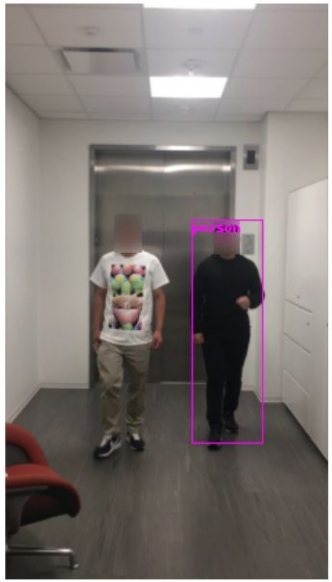
Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. "Adversarial patch." arXiv preprint arXiv:1712.09665 (2017).

Sachin Joglekar, "Adversarial patches for CNNs explained" (2018)

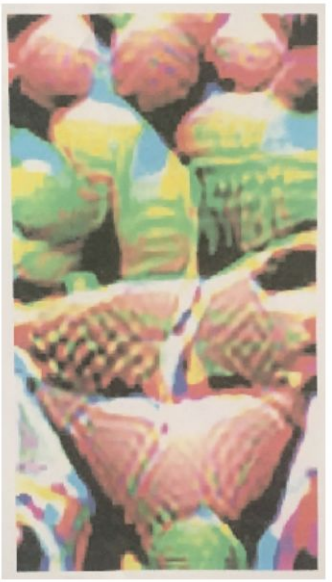


Brown, Tom B., Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. "Adversarial patch." arXiv preprint arXiv:1712.09665 (2017).

Bonus: Adversarial Attacks (eg. T-shirts for Privacy)







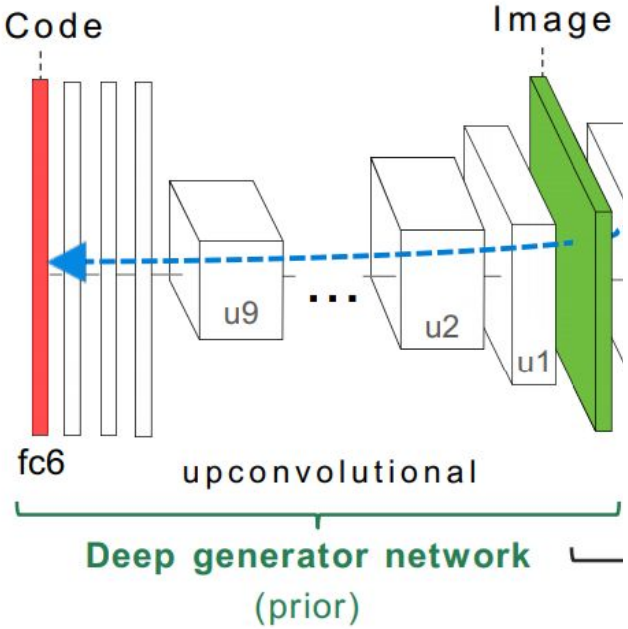
Xu, K., Zhang, G., Liu, S., Fan, Q., Sun, M., Chen, H., ... & Lin, X. (2019). Evading Real-Time Person Detectors by Adversarial T-shirt. arXiv preprint arXiv:1910.11099.

Feature Visualization: Regularizations

	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.				■	

Feature Visualization: Generative prior

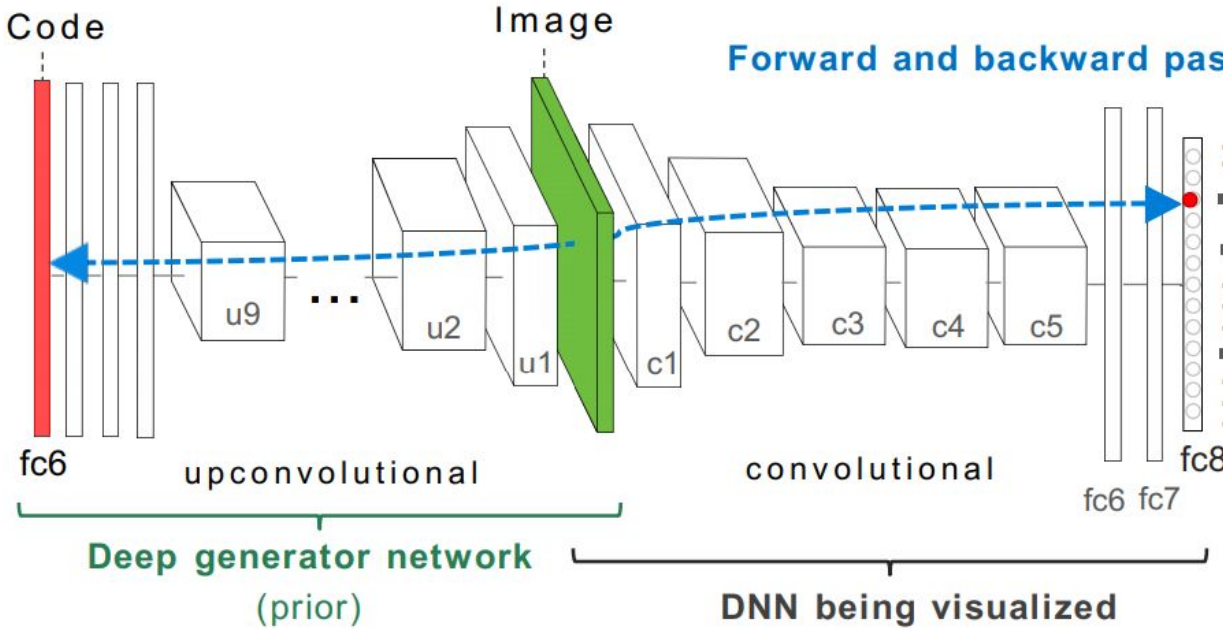
1) A Deep Generator Network (DGN) is trained to invert from features to image.



Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." NIPS 2016.

Feature Visualization: Generative prior

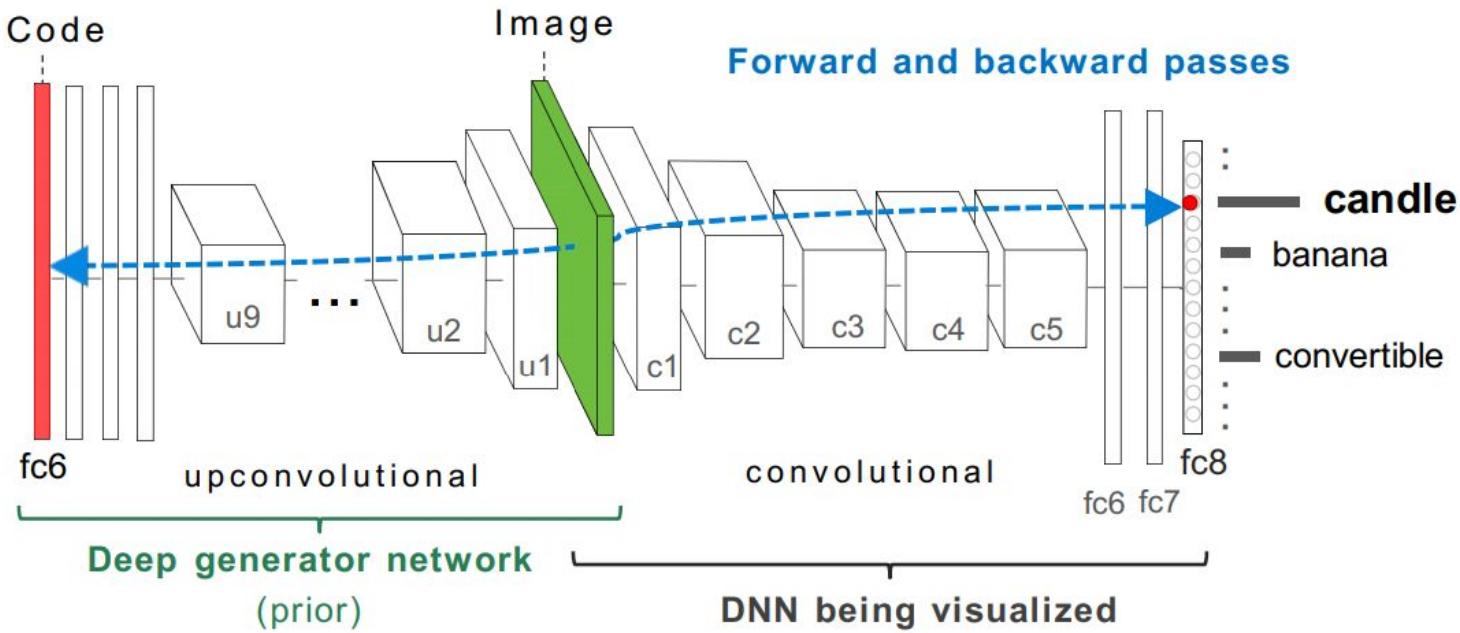
1) A Deep Generator Network (DGN) is trained to invert from features to image.
2) A DNN to visualize is chosen.



Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." NIPS 2016.

Feature Visualization: Generative prior

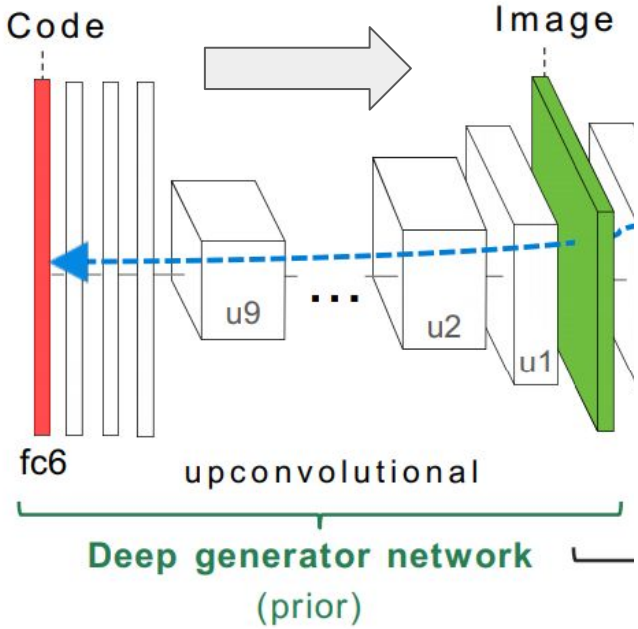
3) Fixing parameters in DGN & DNN, a feature is optimed with backprop to maximize a **class** (eg. candle).



Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." NIPS 2016.

Feature Visualization: Generative prior

4) The image corresponding to the optimzed feature is synthesized with DGN.



Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." NIPS 2016.

Feature Visualization: Generative prior

Nguyen, Anh, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. "Synthesizing the preferred inputs for neurons in neural networks via deep generator networks." NIPS 2016.

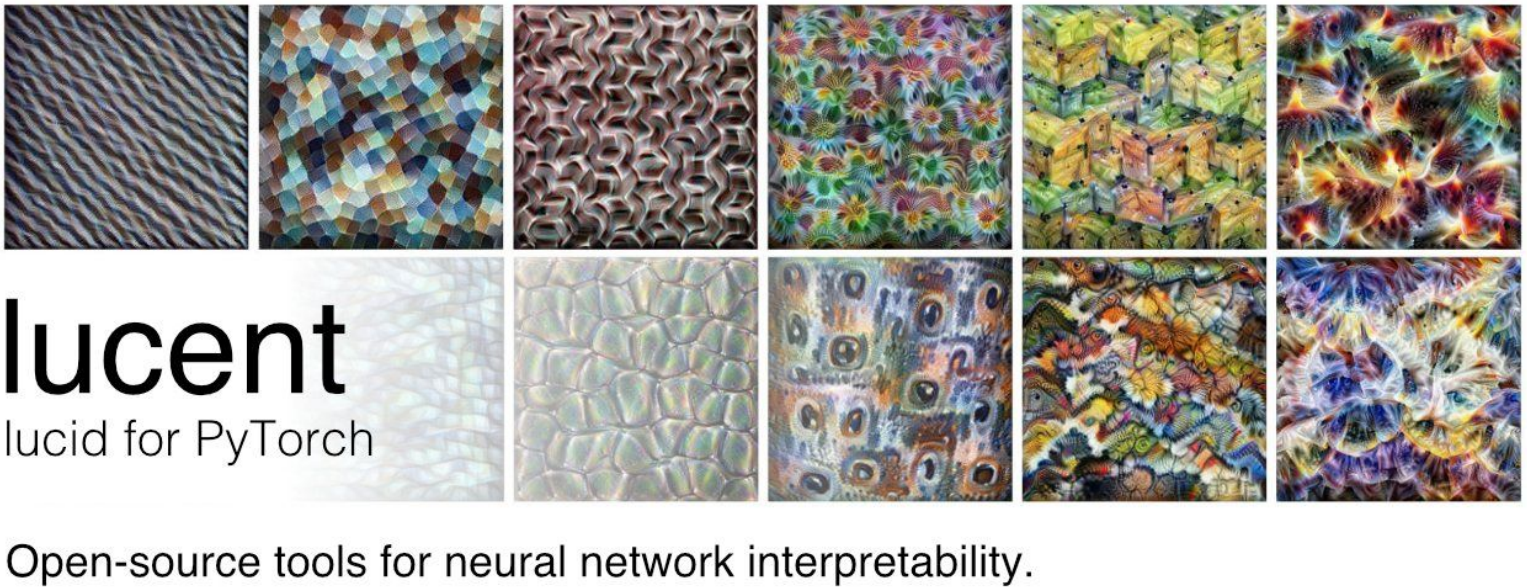
Interpretability for PyTorch

Software: Captum (multiple examples)

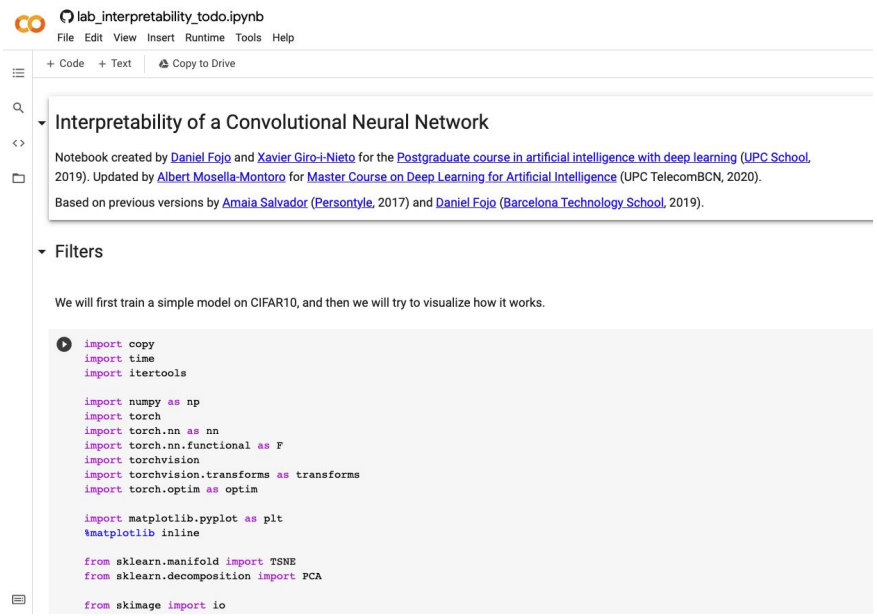
Target Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.99)	pos	1.54	It was a fantastic performance ! pad
pos	pos (0.71)	pos	1.69	Best film ever pad pad pad pad
pos	pos (0.95)	pos	1.49	Such a great show ! pad pad
neg	neg (0.18)	pos	-1.19	It was a horrible movie pad pad
neg	neg (0.22)	pos	-1.70	I 've never watched something as bad
neg	neg (0.38)	pos	-1.29	It is a disgusting movie ! pad

Interpretability for PyTorch

Software: Lucent, based on Kornia [tweet]



(extra) PyTorch Lab on Google Colab



Amaia Salvador
amaia.salvador@upc.edu
PhD 2019
Universitat Politècnica de Catalunya

DL resources from UPC Telecos:

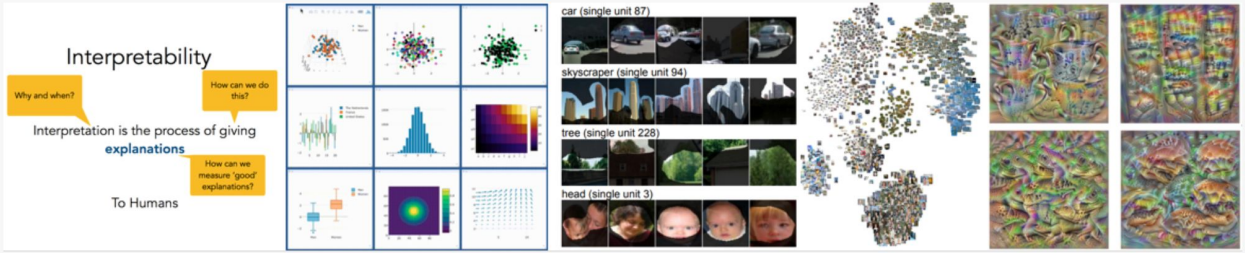
- Lectures (with Slides & Videos)
- Labs

Learn more

ICCV 2019 Tutorial on

Interpretable Machine Learning for Computer Vision

Auditorium, COEX Convention Center, Seoul, Korea
Sunday afternoon, Oct 27, 2019



Speakers

Bolei Zhou
CUHK

Andrea Vedaldi
Oxford

Alexander Binder
SUTD

Alan L. Yuille
JHU

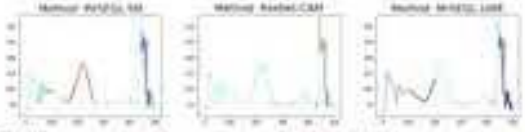
Additional material: Interpretability of Time Series

Why?

- **TSC used in many critical human-related tasks**: human activity recognition, sports sciences, medical applications.

Landing
Classes
-Normal
-Bending
-Stumble

- **Explanations often required**: saliency maps highlighting the important, relevant parts of the TS critical for classifiers to make decision
- **Non-agreement of Explanations** happens when methods provide different saliency maps



Challenge: How to assess and objectively compare TSC explanation methods?

What?

- **Evaluation of Explanations Quality**: simple, informative, interpretable by human users, etc.
- **Informativeness of Explanation**: more informative explanation associates with the higher capability to influence classifiers to correctly identify a class.
- **Devise an evaluation method & a metric** to measure informativeness of different methods

Additional material

Bau, D., Zhu, J. Y., Strobelt, H., Lapedriza, A., Zhou, B., & Torralba, A. Understanding the role of individual units in a deep neural network. PNAS 2020. [tweet]

Oana-Maria Camburu, “Explaining Deep Neural Networks”. Oxford VGG 2020.

Adebayo, Julius, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. "Sanity checks for saliency maps." NeurIPS 2018

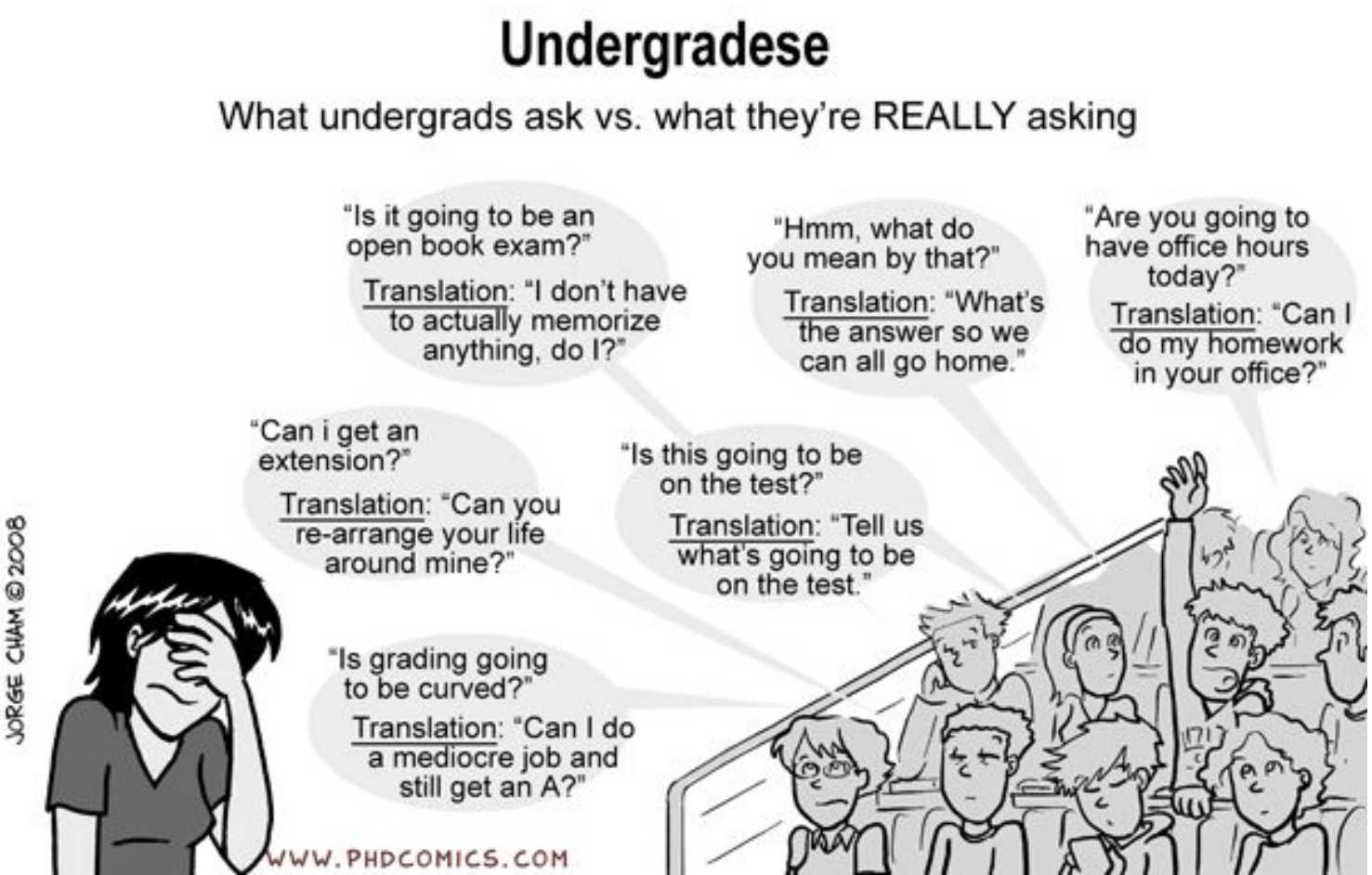
#HINT Selvaraju, Ramprasaath R., Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. "Taking a hint: Leveraging explanations to make vision and language models more grounded." ICCV 2019. [blog].

#Ablation-CAM Ramaswamy, Harish Guruprasad. "Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization." WACV 2020.

Chris Olah et al, “An Overview of Early Vision in InceptionV1”. Distill .pub, 2020.

Rebuffi, S. A., Fong, R., Ji, X., & Vedaldi, A. There and Back Again: Revisiting Backpropagation Saliency Methods. CVPR 2020. [tweet]

Questions ?


Undergradese
What undergrads ask vs. what they're REALLY asking
"Is it going to be an open book exam?"
Translation: "I don't have to actually memorize anything, do I?"
"Hmm, what do you mean by that?"
Translation: "What's the answer so we can all go home."
"Are you going to have office hours today?"
Translation: "Can I do my homework in your office?"
"Can i get an extension?"
Translation: "Can you re-arrange your life around mine?"
"Is this going to be on the test?"
Translation: "Tell us what's going to be on the test."
"Is grading going to be curved?"
Translation: "Can I do a mediocre job and still get an A?"
JORGE CHAM © 2008
WWW.PHDCOMICS.COM