10. DEEP TROUBLES

Stephan Robert-Nicoud
HEIG-VD/HES-SO

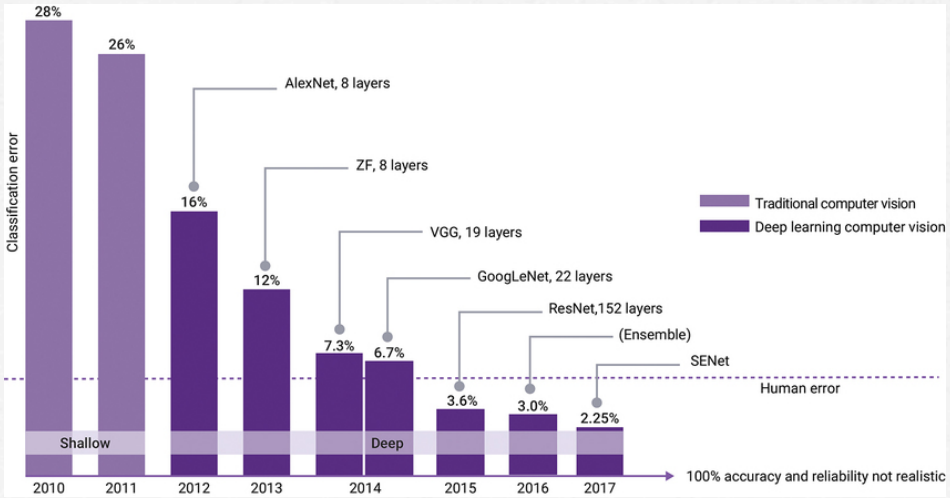
Credit: Andres Perez-Uribe

Objectives

- Understand the limits of Convolutional Neural Networks and the problems arising from the way they are created and the way they work
- Understand the basic principles that can help us visualize the inner working of Convolutional Neural Networks
- Use the visualization techniques to analyze a pre-trained CNN and verify the relevance of the information it is using to solve an object recognition task

CNN advantages (1)

- ILSRVC challenge 2010-2017 —> automatic feature discovery better than "manual" feature extraction
- Before that, the state-of-the-art solutions for object recognition were based on feature extractors like HOG (Histogram of Oriented Gradients) and SIFT (Scale invariant Feature Transform).



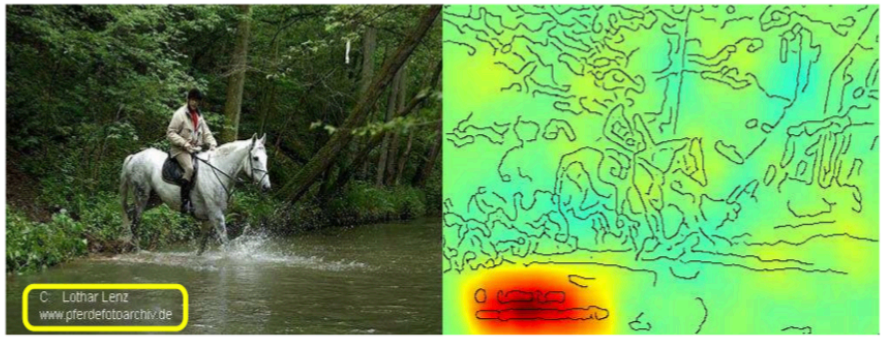
Chihuahua or muffin ?

Astonishingly, a fine-tuned CNN (transfer learning) works very well on the chihuahua vs muffin challenge!

Meaningfulness of solutions

- The CNN is just a very powerful algorithm that
  - computes automatic features from the input data
  - and finds correlations between input patterns to the computed features, and classes
  - Are those features meaningful ?
  - Are those correlations relevant ?

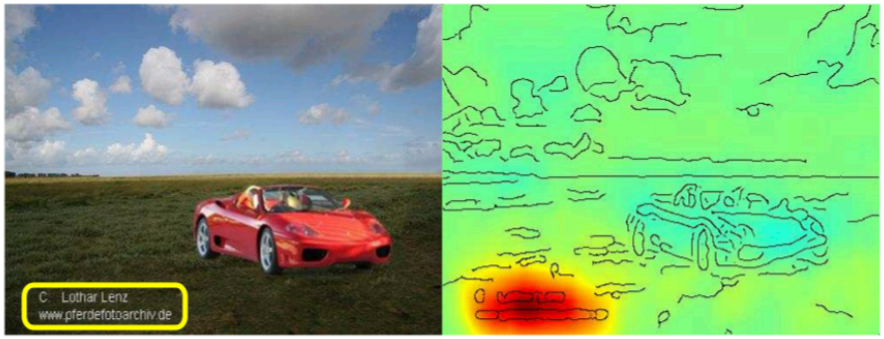
Is there a horse in the image ?

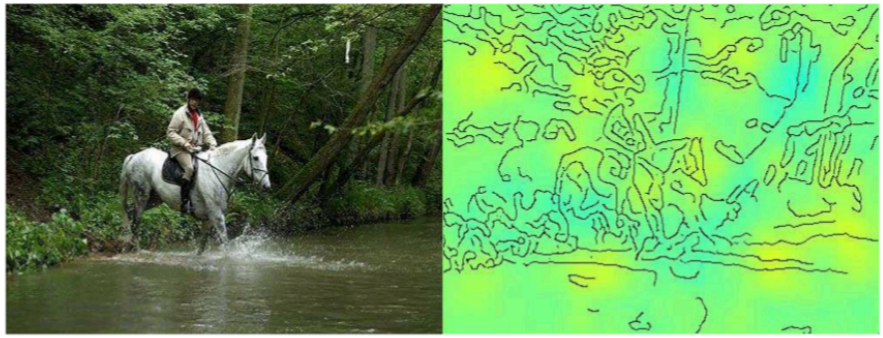


Source tag present

↓

Classified as horse





No source tag present

↓

Not classified as horse

S Lapuschkin et al., 2019

Performance guarantee

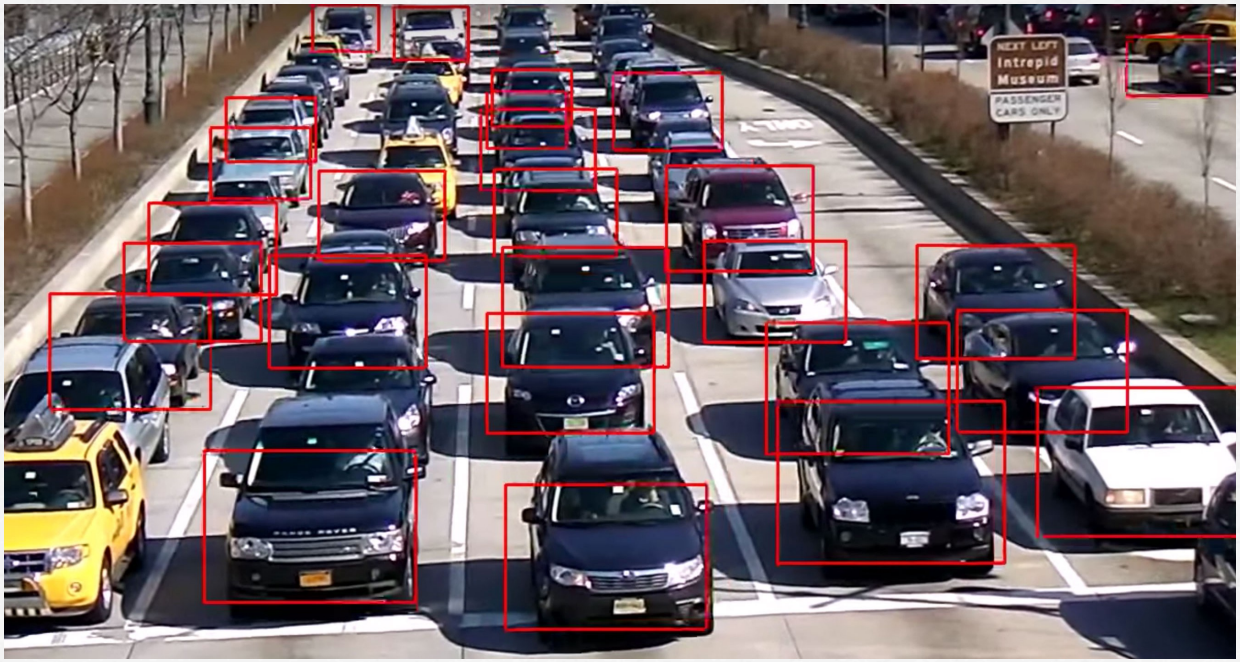
- A good performance on a benchmark dataset is not a guarantee of good performance later in real life
- The truth is that we cannot "for sure" predict the behavior of a CNN-based solution on new inputs

Is there a person here ? (2)

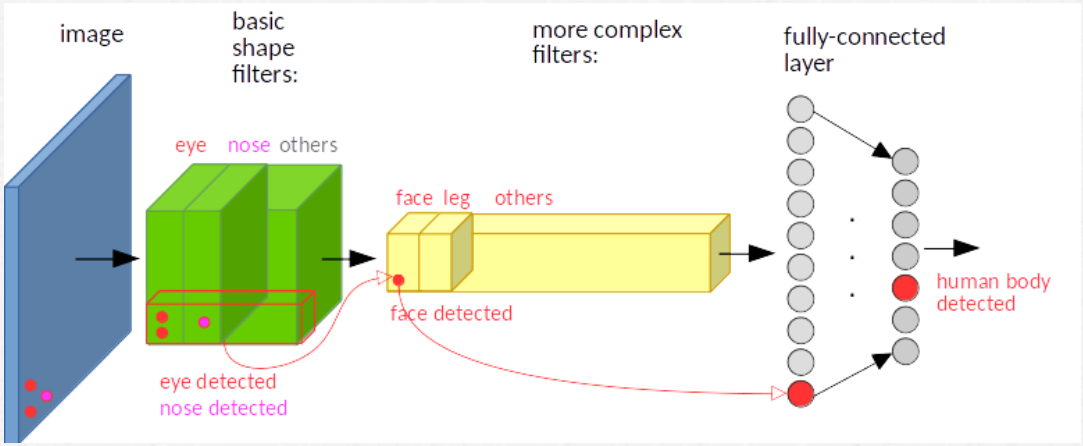
- Biases in the training datasets are then reflected in the applications using the models we created with those data
- There is an urgent need for more inclusive systems (gender, sex, skin color, age, minority, etc) but this requires collective more inclusive data

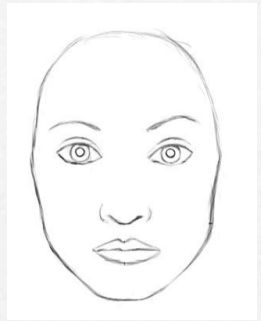
CNN advantages (2)

- Hierarchical feature extraction allows for spatial translation invariance and the recognition of objects appearing at different sizes

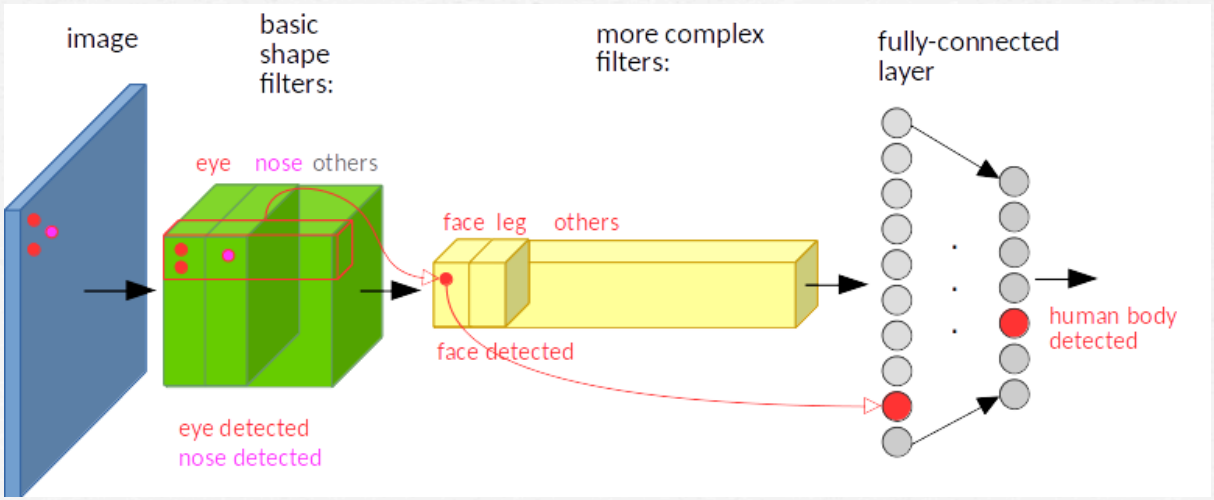


Spatial translation invariance





person: 0.88



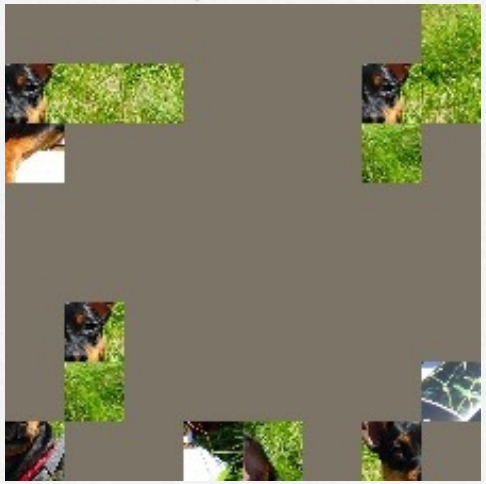
person: 0.85

“Brute force” correlations

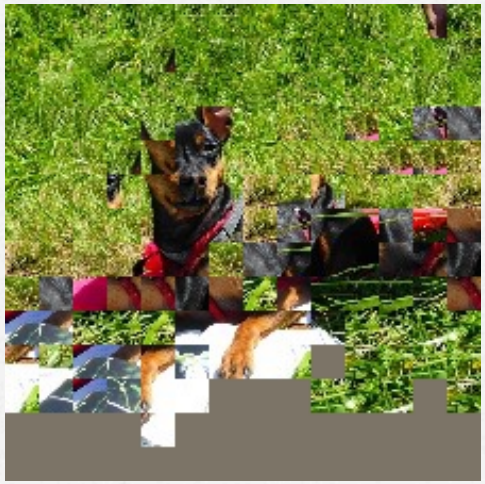
A CNN does not “understand” the data it is processing… it is just detecting and computing features to make a decision. A face with more eyes and appearing anywhere can be associated to the class “face” even more strongly than a normal face.

Output maximization by image occlusion

0.47



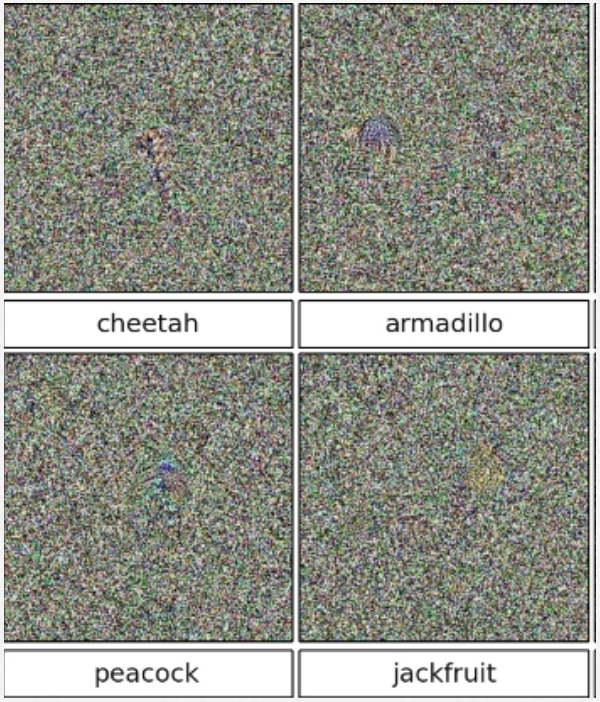
0.97



0.99

pepita = miniature pinscher ? (Satizabal, HEIG-VD 2016)

Adversarial attacks (1)



Nquyen, Yosinski, Clune, 2014

- Since 2013, researchers started to find that certain high-performance CNNs were surprisingly easily fooled.
- Then, they deliberately tried to find a way to systematically fool them (e.g., using ascent gradient or evolutionary algorithms)

Physical-robust attack: a STOP signal is perceived as a speed limit sign (max 45 mph)

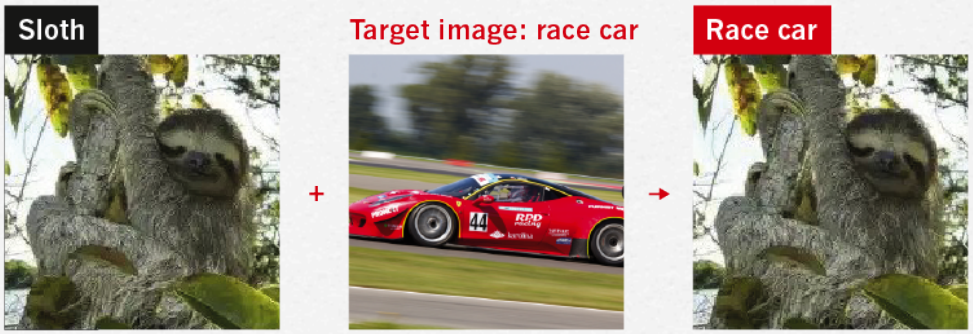
K Eykholt et al, CVPR 2018

Adversarial attacks (2)

PERCEPTION PROBLEMS

Adding carefully crafted noise to a picture can create a new image that people would see as identical, but which a DNN sees as utterly different.

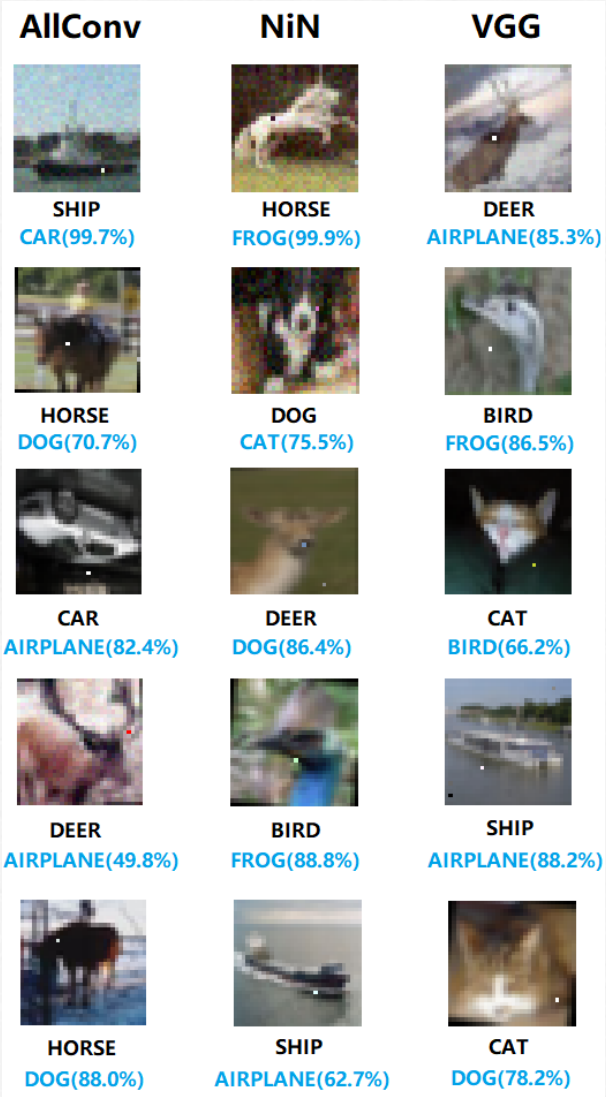
In this way, any starting image can be tweaked so a DNN misclassifies it as any target image a researcher chooses.



©nature

Goodfellow et al, ICLR 2015

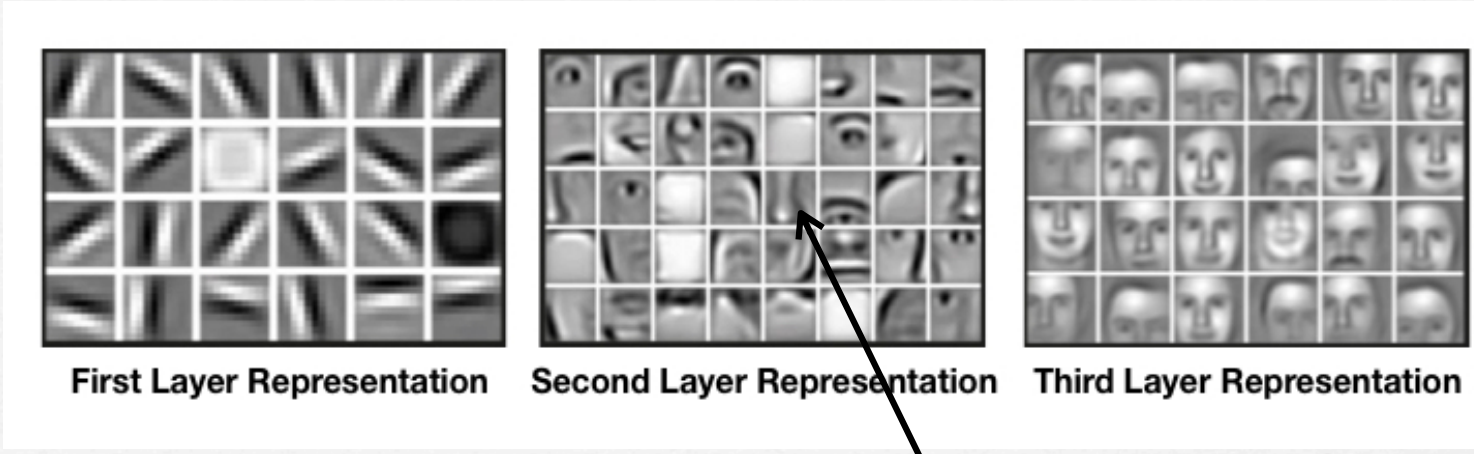
One pixel attack for fooling CNNs



- Jiawei Sun, Danilo Vasconcellos and K. Sakourai from Kyushu University showed that +40% of the ImageNet validation dataset can be perturbed to at least one target class by modifying a single pixel !
- Three CNNs were used for this study: AllConv, NiN and VGG.

arXiv:1710.08864v4 (22.2.18)

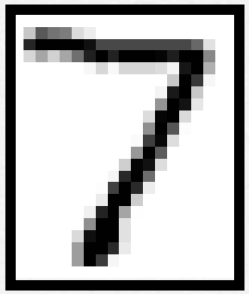
Visualization tools



images that maximize the activation of each filter

- Feature Map visualization
- Activation maximization
- Filter activation statistics
- Occlusion analysis
- Class Activation Maps
- Deconvolution

Feature Map visualization



input (28x28)

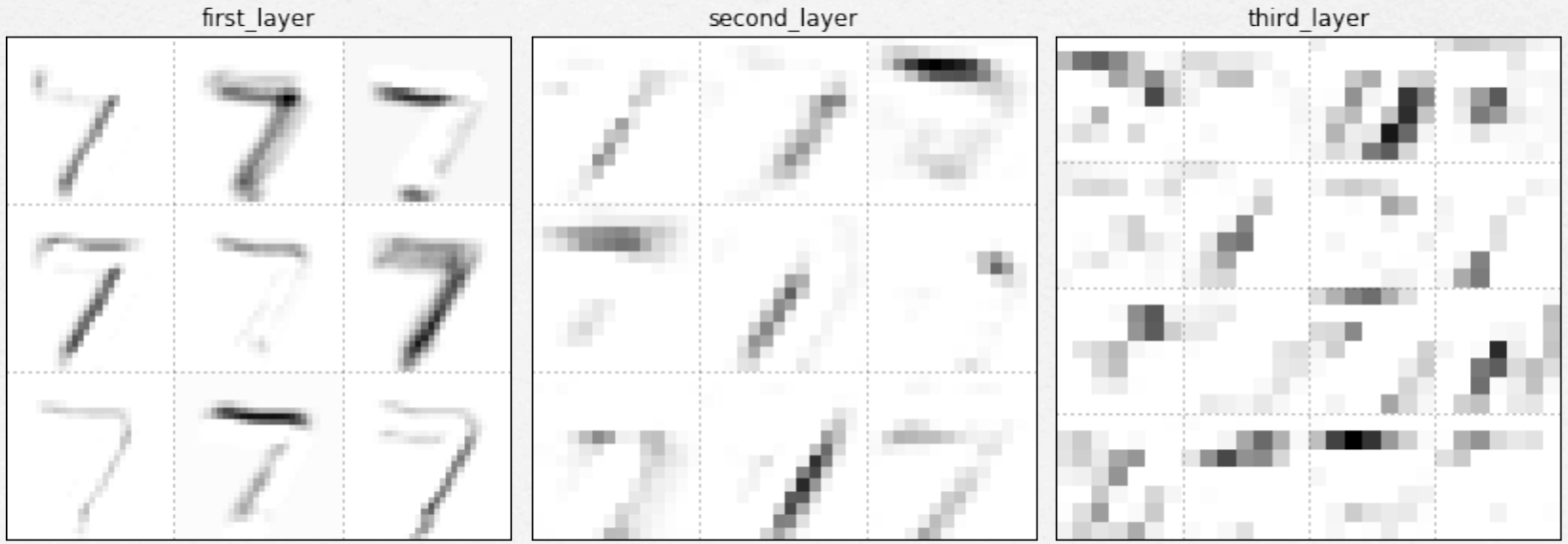
9x[5x5]

maxpool

9x[5x5]

maxpool

16x[3x3]



28x28 14x14 7x7

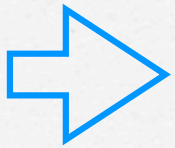
Outputs of the convolutional filters

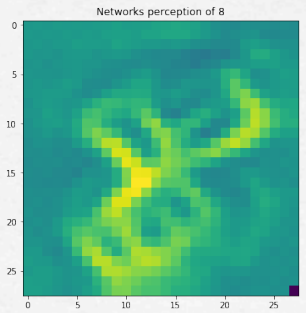
Activation Maximization

- The output of the first convolutional layer is easy to interpret. Let's simply visualize it as an image.
- Subsequent convolutional filters operate over the outputs of previous filters (which indicate the presence or absence of some "templates"), making them hard to interpret.
- Idea: what sort of input pattern maximizes the activation of a particular filter ?
- Use ∂ Activation / ∂ input to modify a random input image and maximize the activation of a given output. Example:

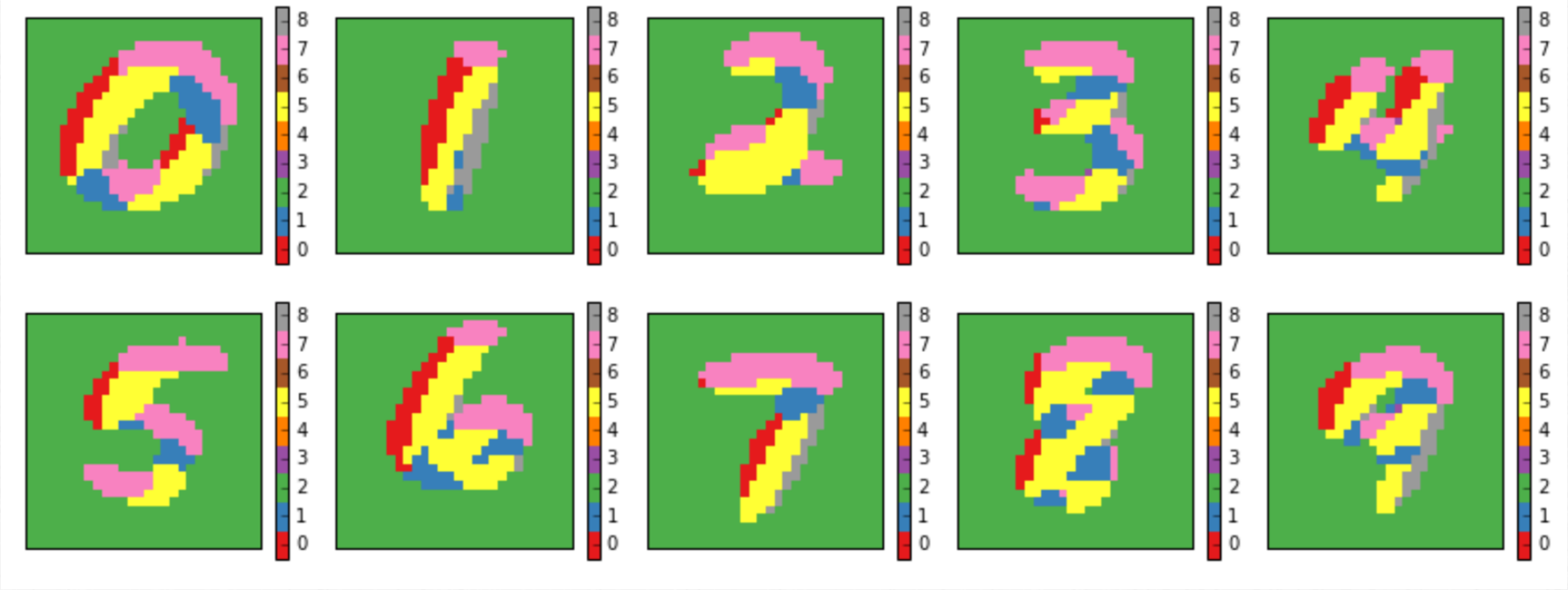
28x28

random image





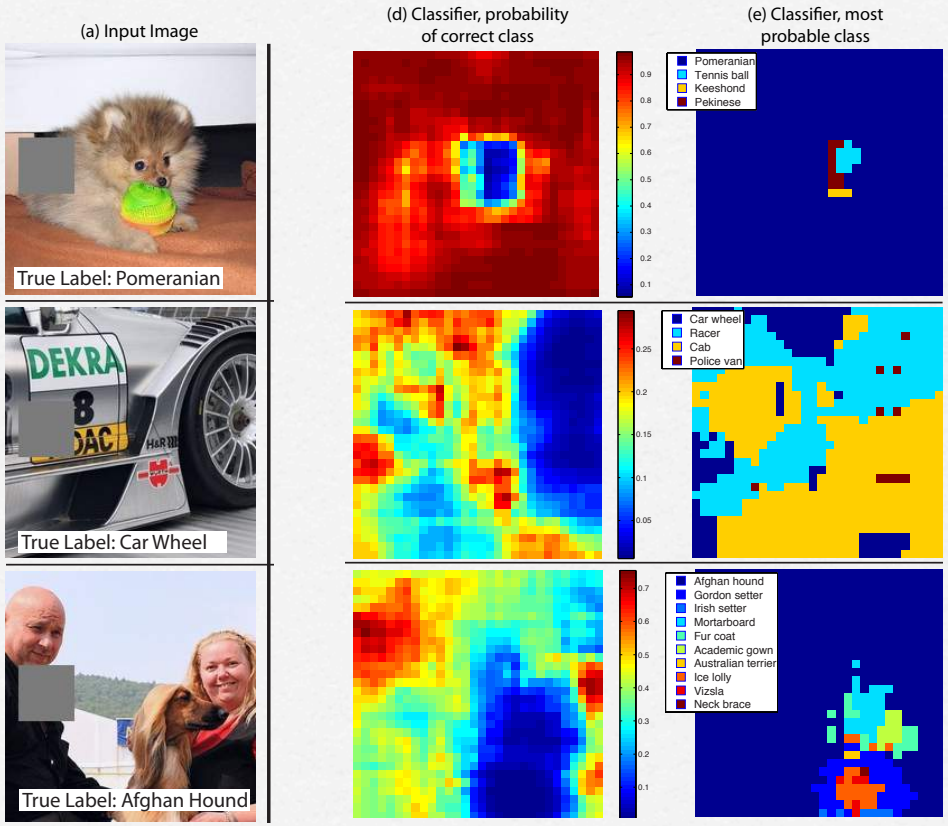
Filter Activation Statistics



The red pixels indicate that for them, the kernel 0 is the most frequent filter (for all same digits in the database) with the highest activation

Occlusion Analysis

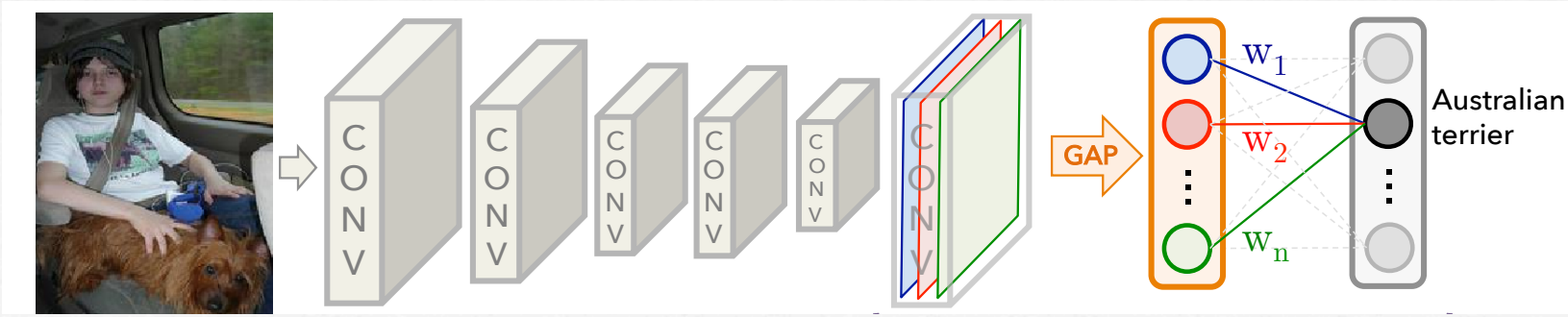
(Zeiler & Fergus, 2013)



What is the most discriminative object or part of the image that lets the CNN decide what is the label of a given image ?

Class Activation Maps (1)

- A Global Average Pooling (GAP) layer computes the mean of the activations of the filters of a given layer.
- We replace the fully-connected part by a GAP layer followed by a dense linear layer using a softmax output. We train this network and obtain weights w_1 to w_n.



from arXiv:1512.04150v1

Class Activation Maps (2)

- The CAM for a given image is the weighted sum of the last-layer's filter activations. It is finally upsampled to match the input image size.
- The result is a "heat-map" indicating what portion of the image the CNN is paying more attention.

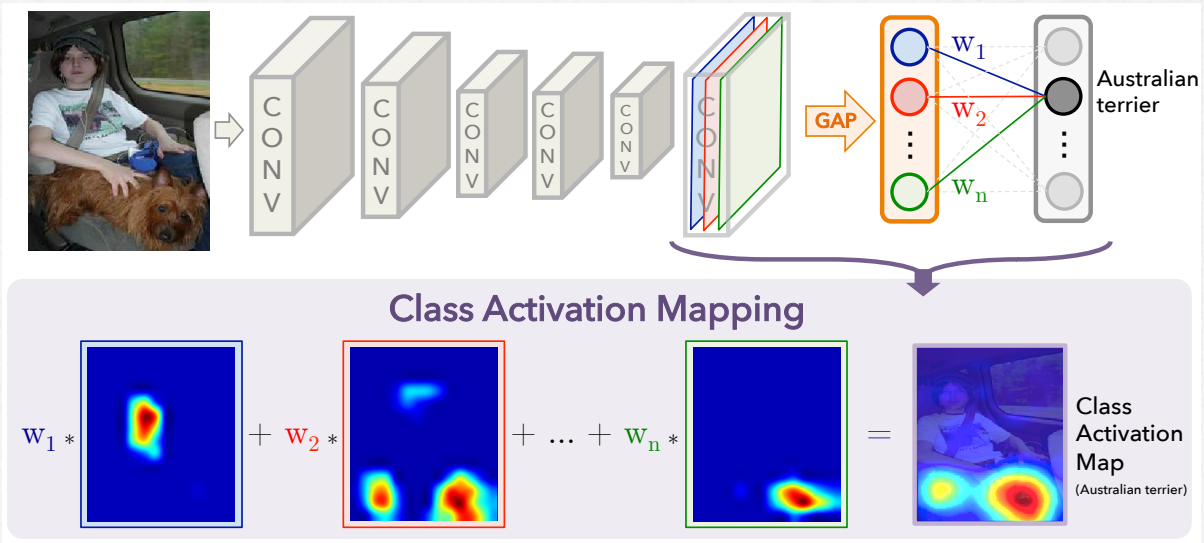
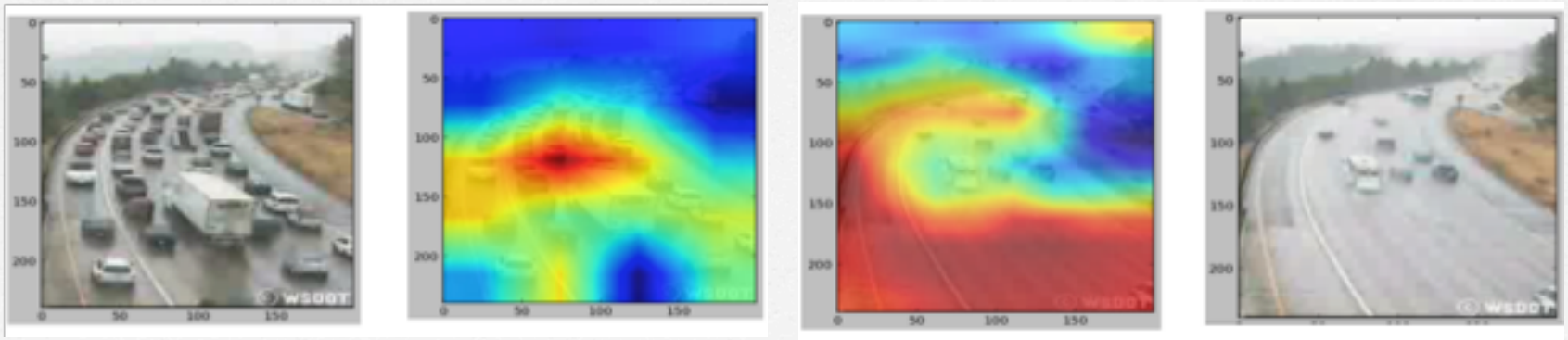


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Example: traffic density estimation

- traffic density

Examples of low, medium and high traffic



presence of many vehicles = high traffic / absence of vehicles = low traffic

Microscope OpenAI

ResNet v2 50

ResNets use skip connections to enable stronger gradients in much deeper networks. This variant has 50 layers.

ResNet v2 50

1,001 predictions/Softmax
1,001 1,1 logits/Conv2D
2,048 block4/unit_3/add
3,3 block4/unit_3/conv2...
1,1
2,048 block4/unit_2/add
3,3
1,1
2,048 block4/unit_1/add
3,3
1,1
1,024 block3/unit_6/add
3,3
1,1
1,024 block3/unit_5/add
3,3
1,1
1,024 block3/unit_4/add
3,3
1,1
1,024 block3/unit_3/add
3,3

block4/unit_3/conv2/Relu

Type: Relu
Channels: 512
Convolution: [3,3]

Technique

- Feature Visualization
- DeepDream
- Dataset Samples
- Caricature
- Text Feature Visualization

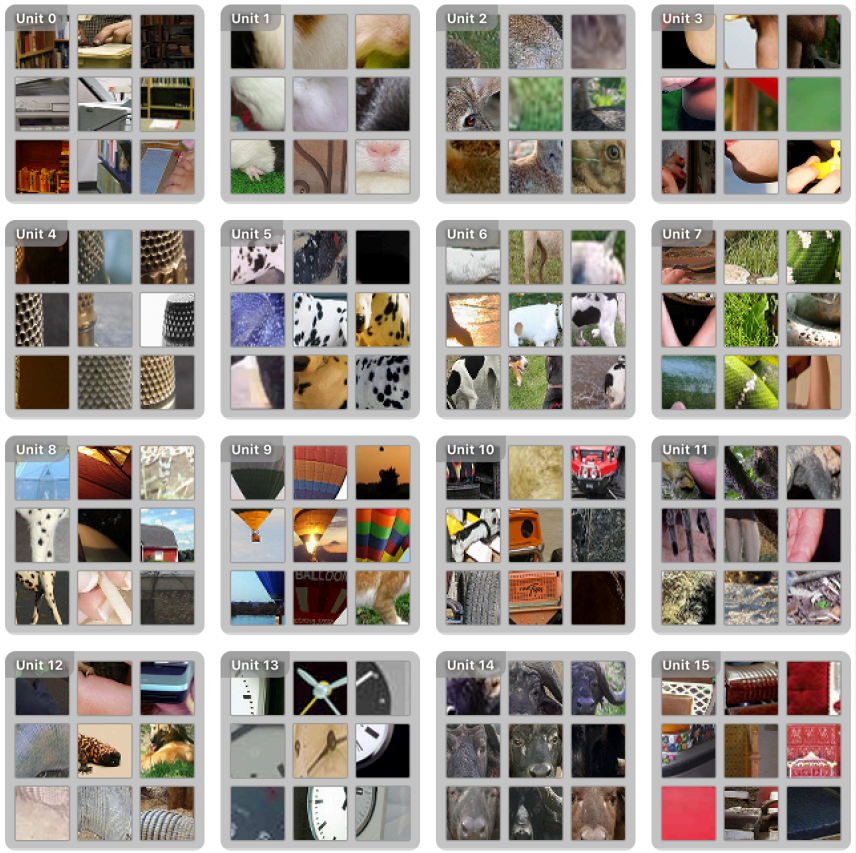
Pieces of images from the training dataset that result in the largest activations from the given unit.

View

Image Size

Resize Behavior

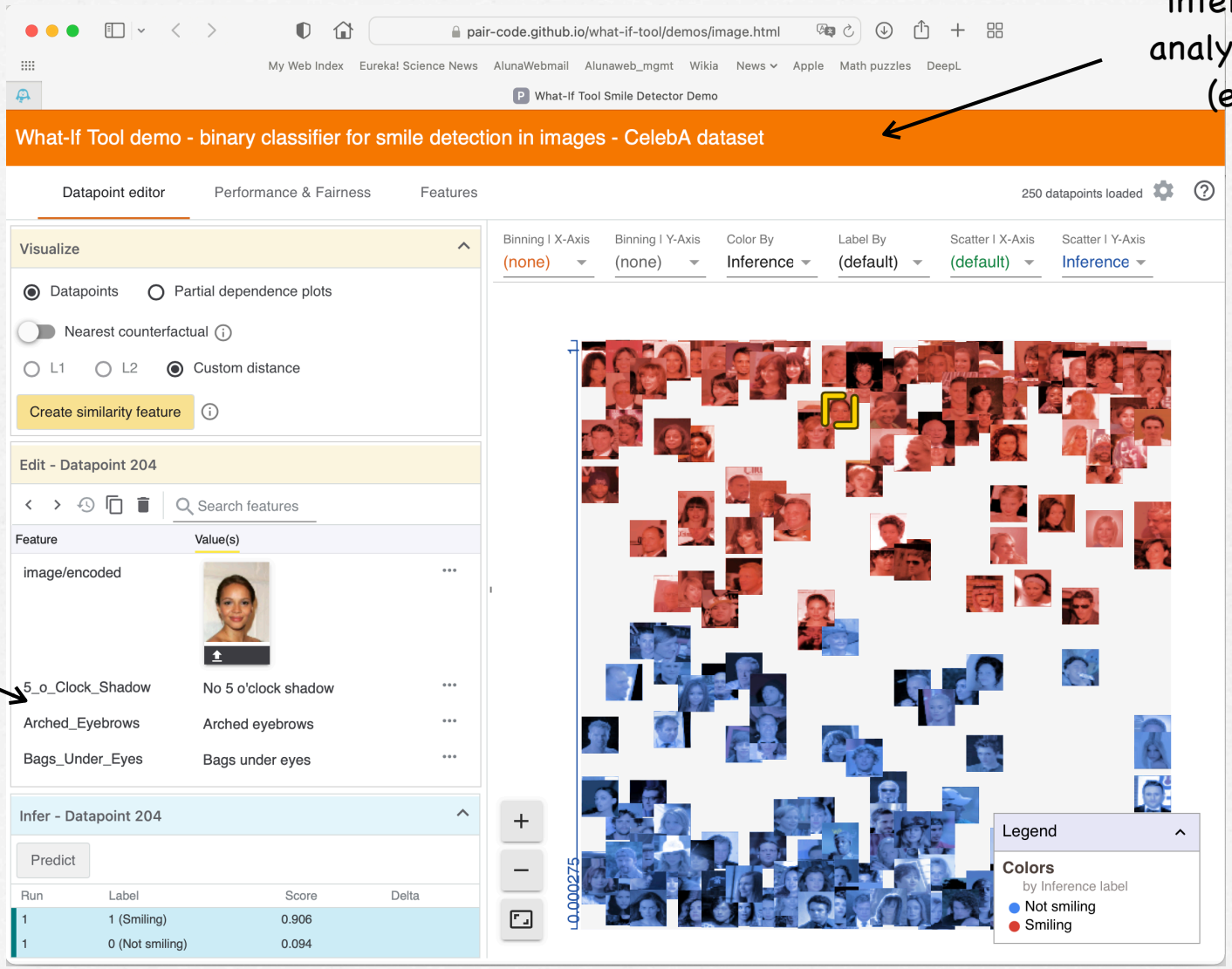
- Crop image
- Scale image



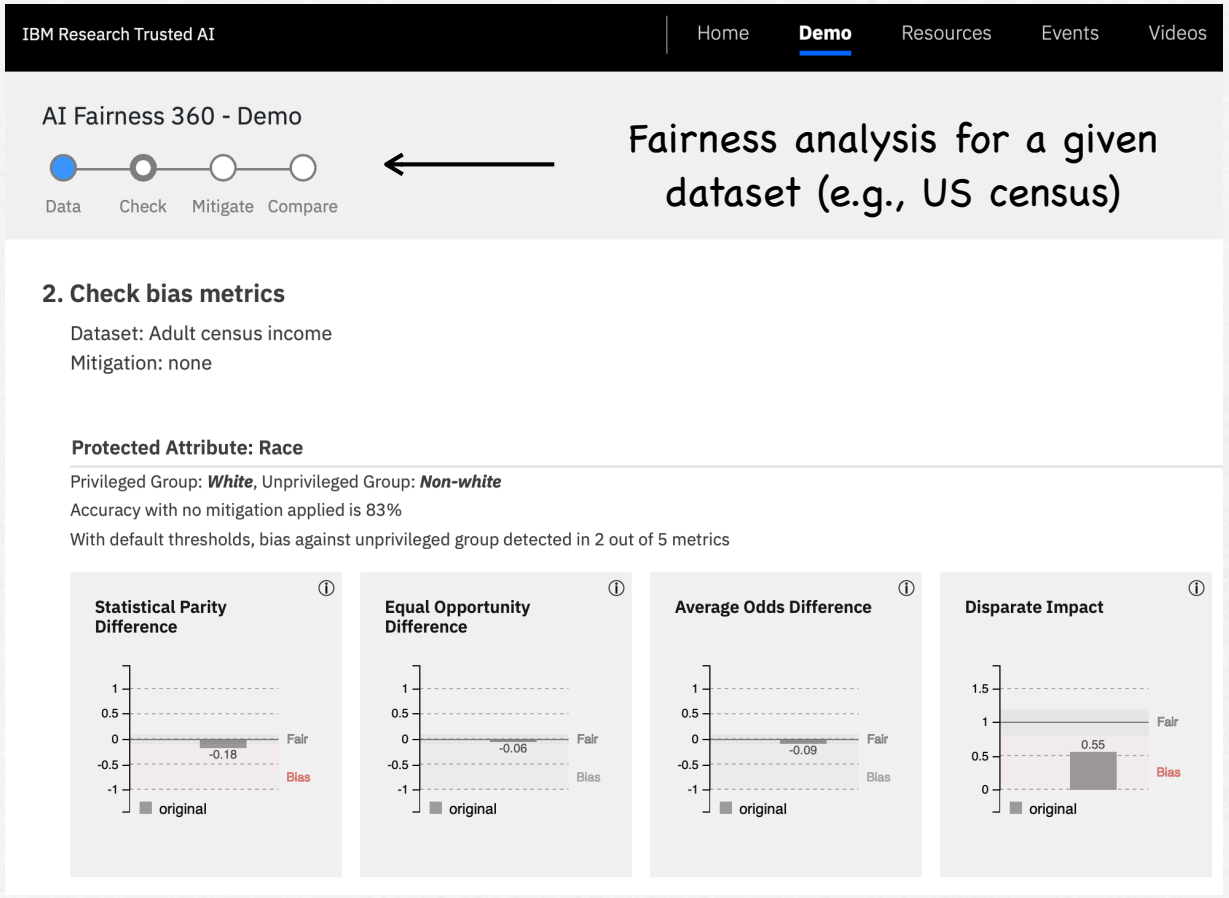
Google's What-If tool

interactive tool to analyze models/data (e.g., CelebA)

human understandable features



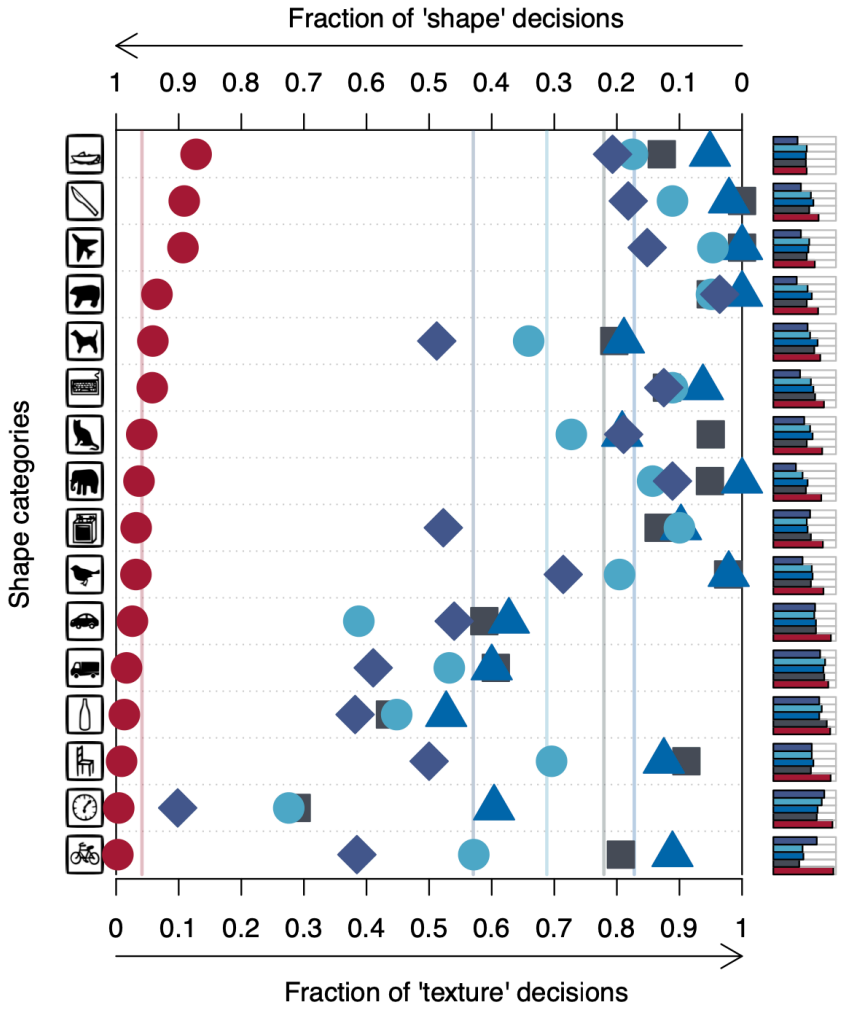
IBM's AI fairness tool



Idea: can we predict gender, sex or skin color from the input data ?

Biases of ImageNet-trained CNNs

Figure 4: Classification results for human observers (red circles) and ImageNet-trained networks AlexNet (purple diamonds), VGG-16 (blue triangles), GoogLeNet (turquoise circles) and ResNet-50 (grey squares). Shape vs. texture biases for stimuli with cue conflict (sorted by human shape bias). Within the responses that corresponded to either the correct texture or correct shape category, the fractions of texture and shape decisions are depicted in the main plot (averages visualised by vertical lines). On the right side, small barplots display the proportion of correct decisions (either texture or shape correctly recognised) as a fraction of all trials. Similar results for ResNet-152, DenseNet-121 and Squeezenet1_1 are reported in the Appendix, Figure 13.



arXiv:1811.12231v3 [ICLR 2019]

Shape vs texture features

Original image

Stylised ImageNet (SIN) dataset

- The SIN dataset is a much harder task than ImageNet (IN)
- A model trained on SIN generalises well on IN, but a model trained on IN does not generalize well on SIN.
- Current CNNs are biased towards learning texture-based features.

Geographical bias of ImageNet

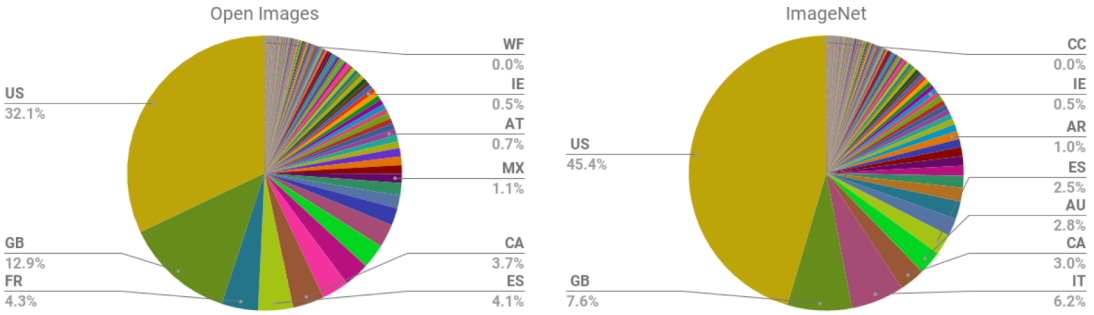


Figure 1: Fraction of Open Images and ImageNet images from each country. In both data sets, top represented locations include the US and Great Britain. Countries are represented by their two-letter ISO country codes.

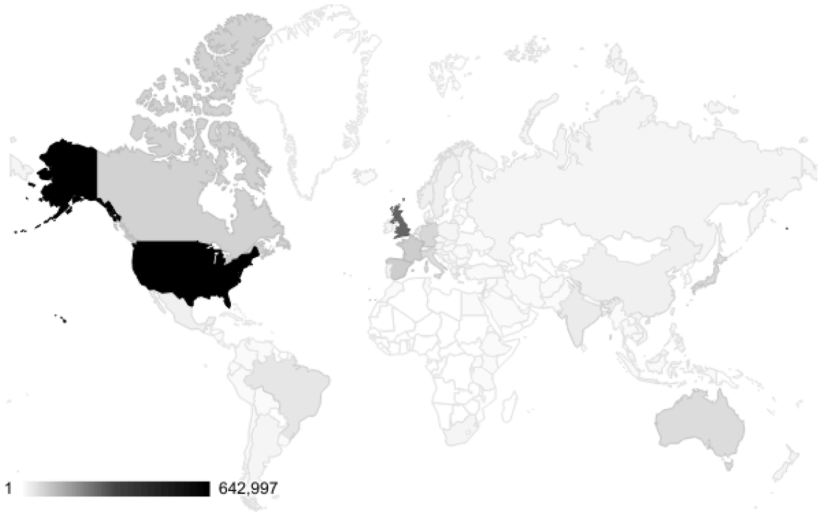


Figure 2: Distribution of the geographically identifiable images in the Open Images data set, by country. Almost a third of the data in our sample was US-based, and 60% of the data was from the six most represented countries across North America and Europe.

present absent

HE IG VD

XAI: eXplainable AI

- Explainable AI (XAI) refers to methods and techniques in the application of artificial intelligence technology (AI) such that human experts can understand the results of the solution.
- If I do not understand the models, how do I trust model predictions ?