Deep Learning Crash Course

Hui Xue

Fall 2021

www.deeplearningcrashcourse.org

National Heart, Lung, and Blood Institute

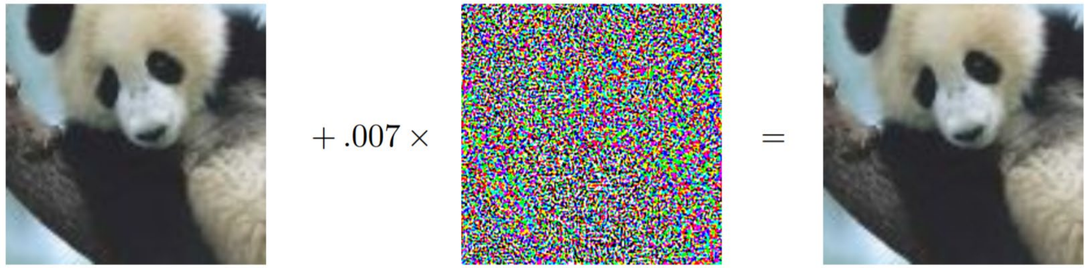
Outline

- Adversarial examples and adversarial training
- Visualization of deep neural networks

Neural networks can be hacked

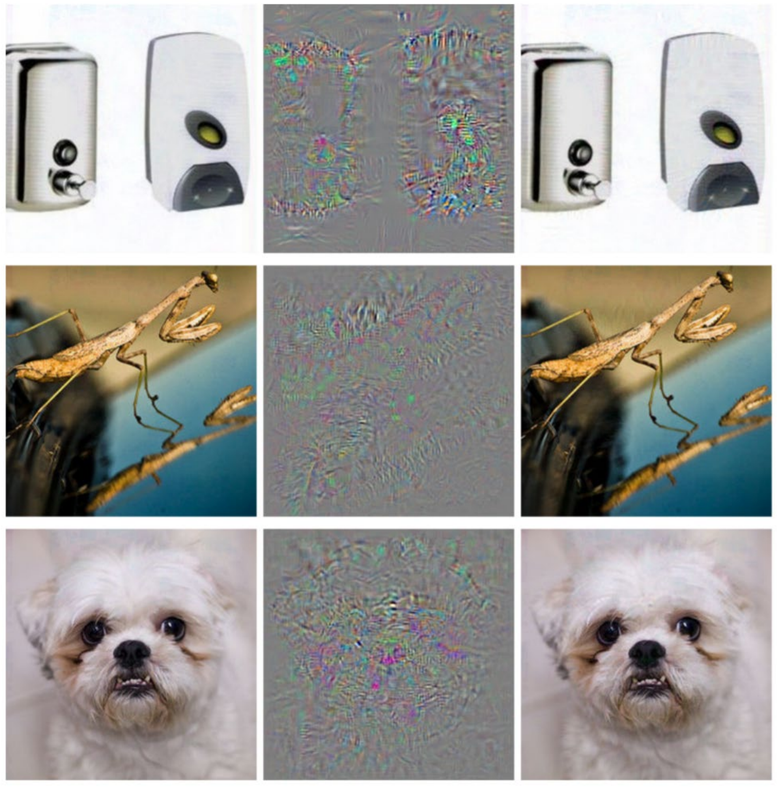
Right after the taking-off of deep learning in 2012

Image + small perturbation → incorrect classification



Panda
57.7%
confidence

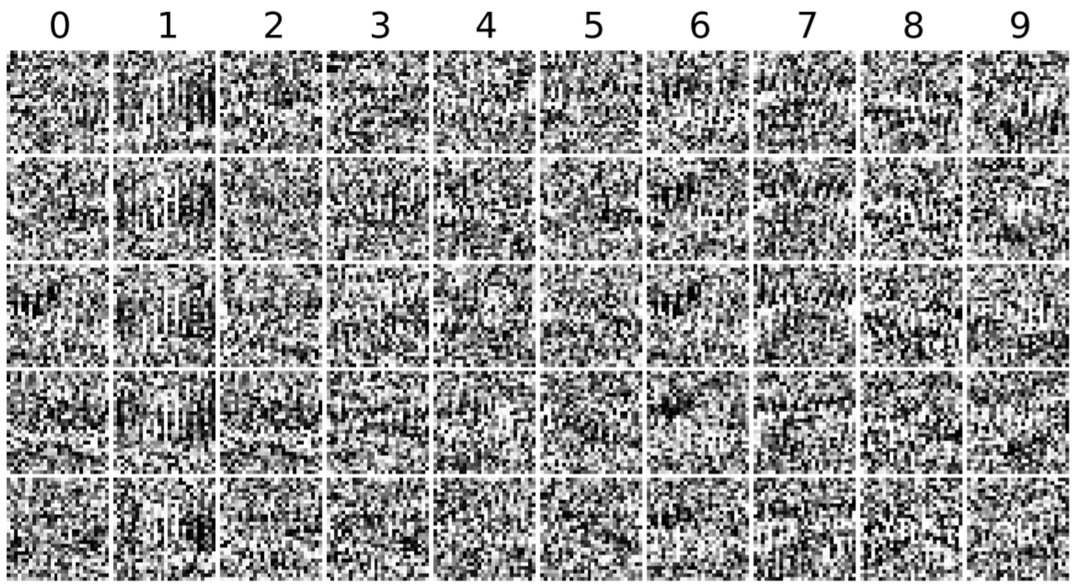
Gibbon
99.3%
confidence

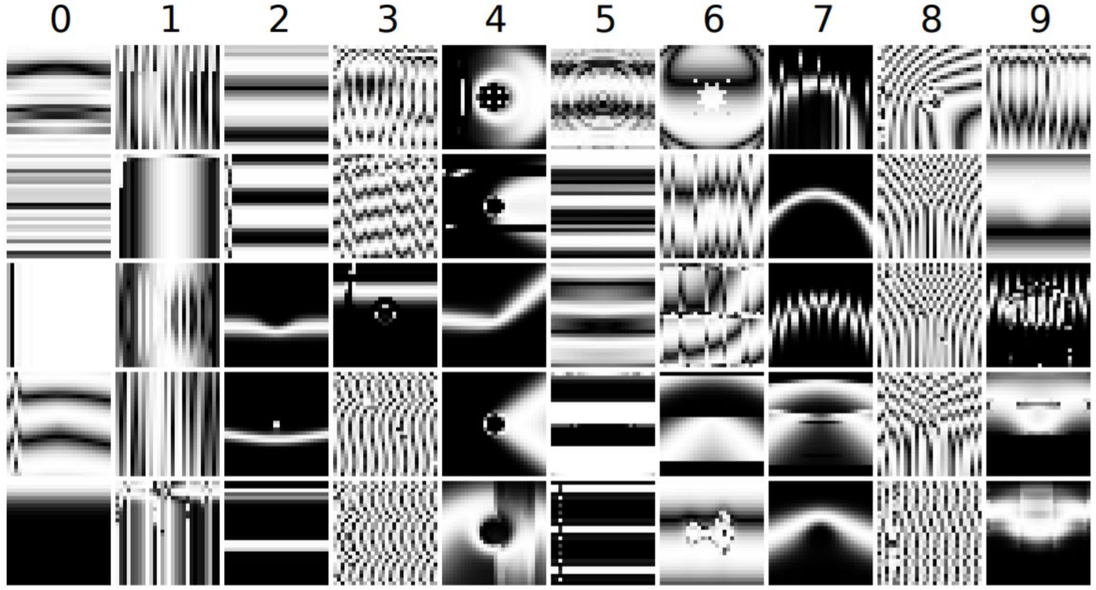


Intriguing properties of neural networks.2014. https://arxiv.org/abs/1312.6199

Explaining and Harnessing Adversarial Examples. 2015. https://arxiv.org/abs/1412.6572

Negative examples





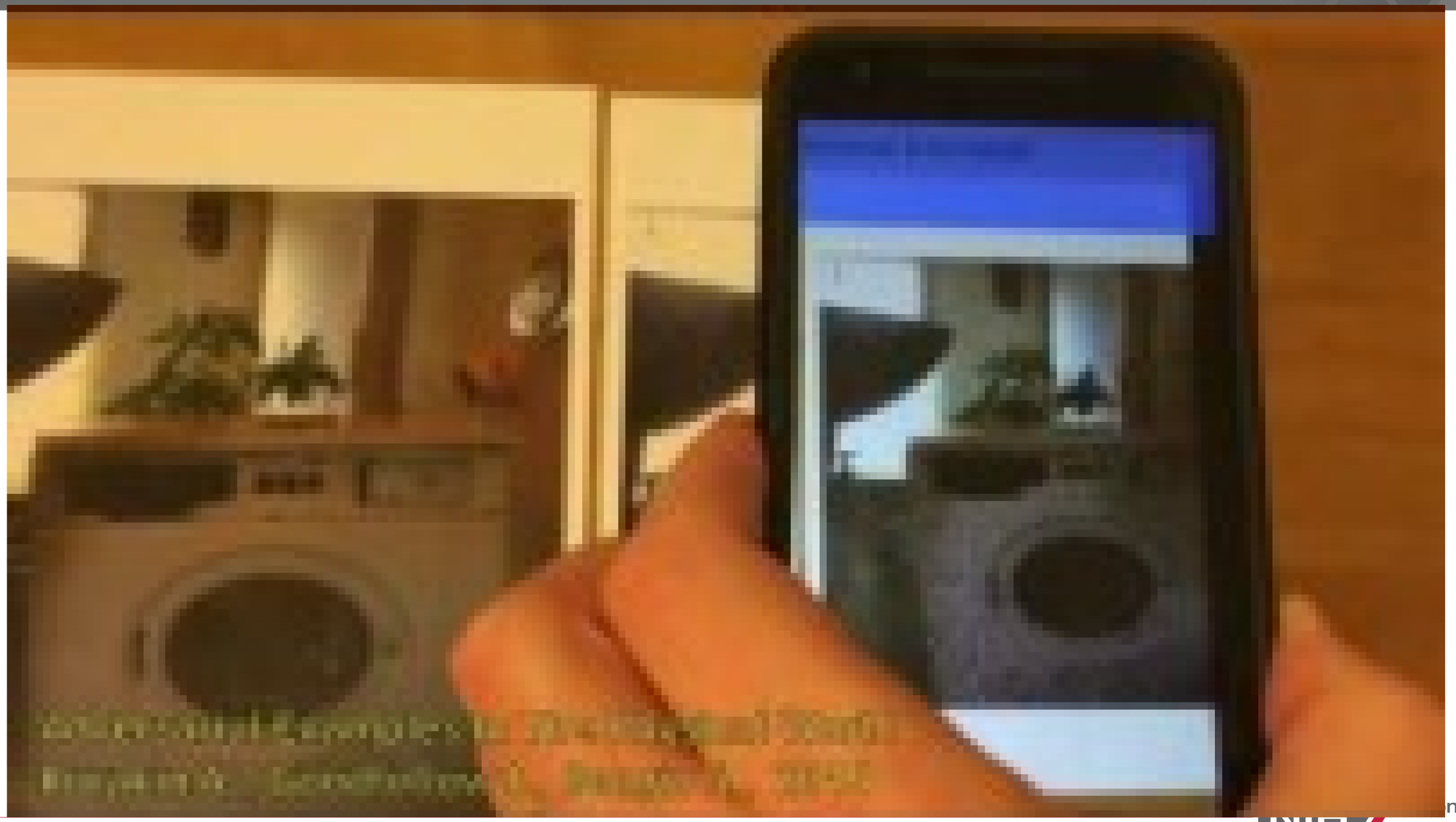
"it is easy to produce images that are completely unrecognizable to humans, but that state-of-the-art DNNs believe to be recognizable objects with 99.99% confidence"

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. 2015. https://arxiv.org/abs/1412.1897

Physical examples

ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. 2019. https://arxiv.org/abs/1804.05810

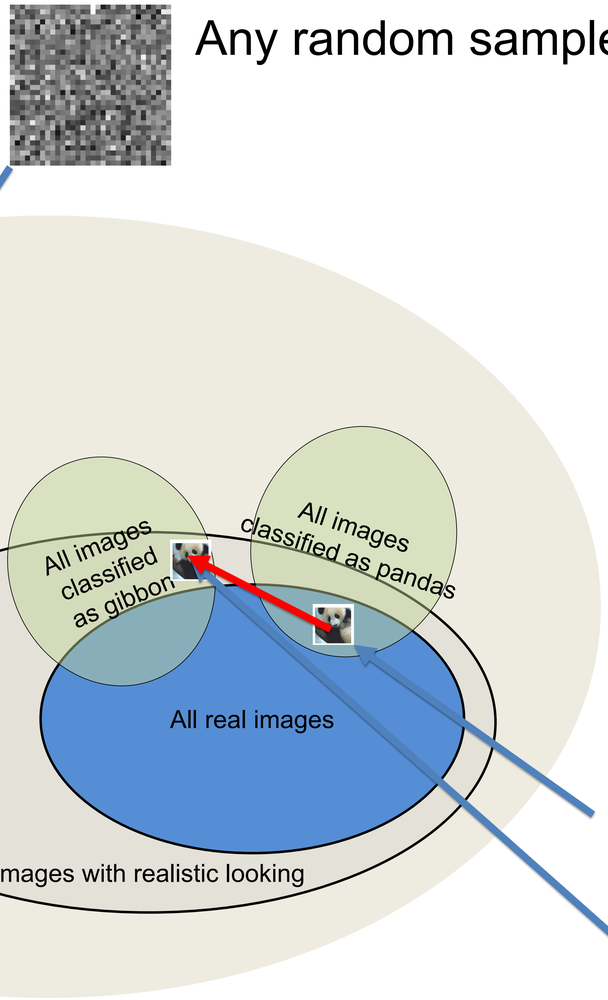
Physical examples



https://youtu.be/zQ_uMenoBCk

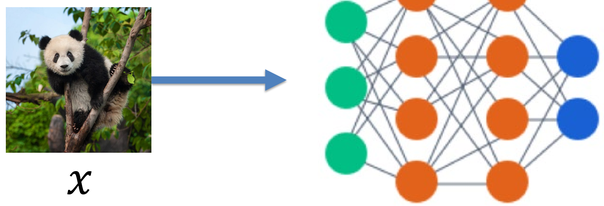
Adversarial examples in the physical world. 2017. https://arxiv.org/abs/1607.02533

Intuition



- Model fits the very limited data manifold in vast sample space
- Model behavior can change drastically beyond the data manifold
- Key is to find the perturbation

Create adversarial examples



x

probability

$$y = \begin{bmatrix} cat & 0.0011 \\ panda & 0.6 \\ gibbon & 0.002 \\ \dots & \dots \end{bmatrix}$$

If we want the model to output a specific class, e.g. gibbon

$$\hat{y} = \begin{bmatrix} cat & 0 \\ panda & 0 \\ gibbon & 1.0 \\ \dots & \dots \end{bmatrix}$$

Define loss between y and $\hat{y}$, e.g. the cross-entropy loss

$$\ell_{ce}(gibbon, y, \hat{y}) = \frac{1}{B}\sum_{i=0}^{B-1} \log(y_{gibbon})$$

Since the target is to change image,

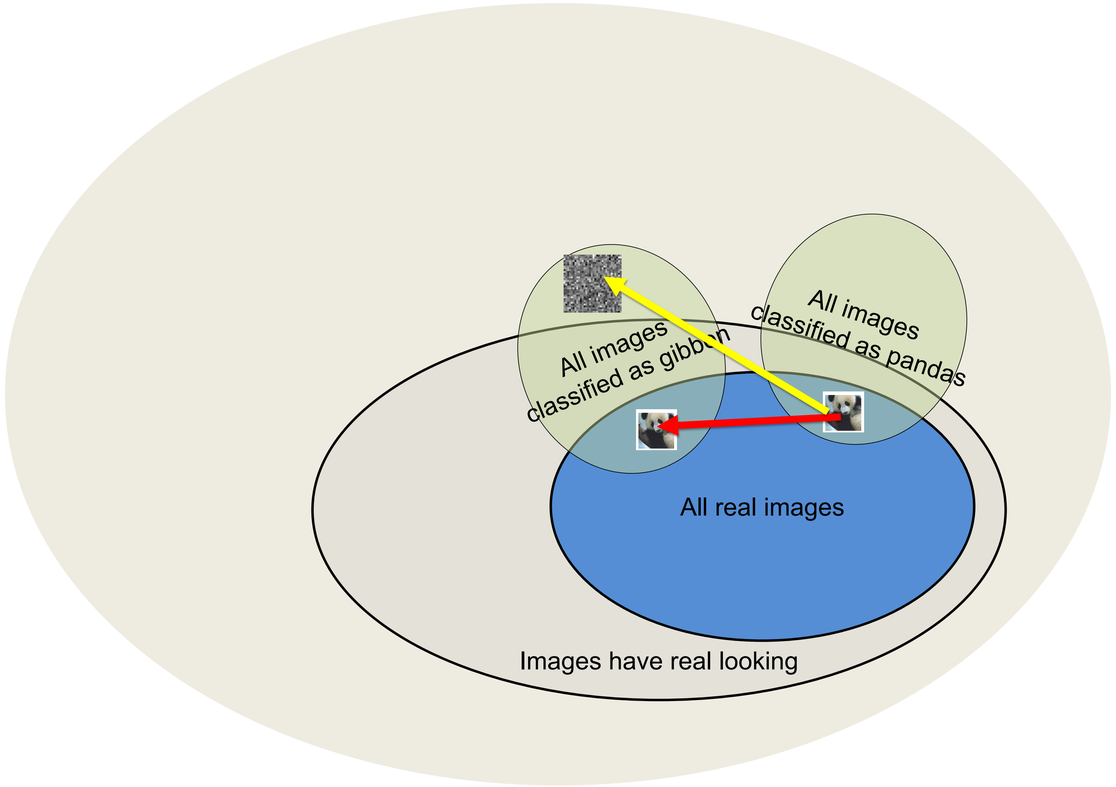
$$x = x + \alpha \frac{\partial \ell_{ce}}{\partial x}$$

https://www.rd.com/wp-content/uploads/2020/03/GettyImages-1060486568.jpg

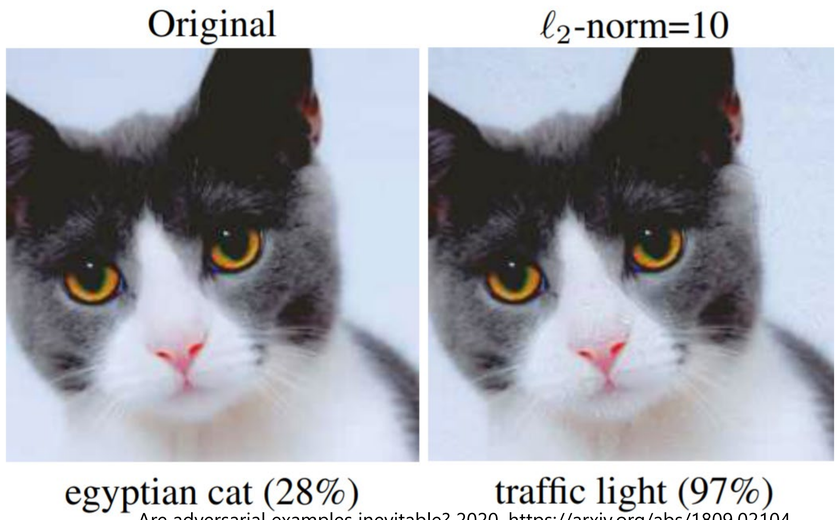
Create adversarial examples for visual assurance

$$x = x + \alpha \frac{\partial \ell_{ce}}{\partial x}$$

Update the image until the model outputs required class classification



- No guarantee to get real looking images
- May not be close to a panda image
- Consider you are the attacker …



Are adversarial examples inevitable? 2020. https://arxiv.org/abs/1809.02104

Iterative method to create adversarial examples

Define loss between y and $\hat{y}$, e.g. the cross-entropy loss

$$\ell_{ce}(gibbon, y(x), \hat{y}) = \frac{1}{B}\sum_{i=0}^{B-1} \log(y_{gibbon})$$

$$\ell(x, x_{ori}, \hat{y}) = \ell_{ce}(gibbon, y(x), \hat{y}) - \ell_{reg}(x)$$

Since the target is to change image,

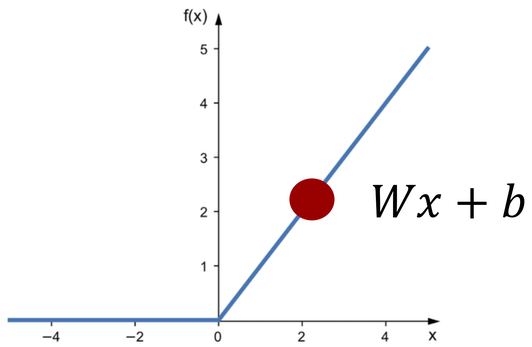
$$x = x + \alpha \frac{\partial \ell_{ce}}{\partial x}$$

- Pick the target class
- Start from the original image
- Perform gradient update step in x until network is fooled

Add the regularization to encourage adversarial sample to have close visual appearance than original sample

$$\ell_{reg}(x) = \lambda \|x - x_{ori}\|_2^2$$

Intuition from simple model: dimensionality



$$y = ReLU(Wx + b)$$

$$\frac{\partial y}{\partial x} = W$$

$$x_{adv} = x + \varepsilon W$$

$$\|x - x_{adv}\|_2^2 = \varepsilon^2 \|W\|_2^2$$

- Create adversarial sample is to move along the direction of derivative
- The amount of change in sample x is related to the norm of weights
- If the network has lots of dimensions, even very small multiplier ε can make large perturbation
- With the number of dimension increasing, $\left\|\frac{\partial \ell}{\partial x}\right\|_2^2 \propto \|W\|_2^2$, the loss is more sensitive to small perturbation in x with higher dimensions

Fast gradient sign method

What is the direction to cause the largest amount of change in loss, given an input x?

$$\frac{\partial \ell(x;\boldsymbol{\theta})}{\partial x}$$ The gradient of loss to input x

If we want to maximize the change in loss, and **also** we want to limit the perturbation in input image :

$$\ell(x_{adv};\boldsymbol{\theta}) = \ell(x;\boldsymbol{\theta}) + (x_{adv} - x)^{\boldsymbol{T}}\frac{\partial \ell(x;\boldsymbol{\theta})}{\partial x}$$

Subject to :

$$\|x - x_{adv}\|_\infty < \varepsilon$$

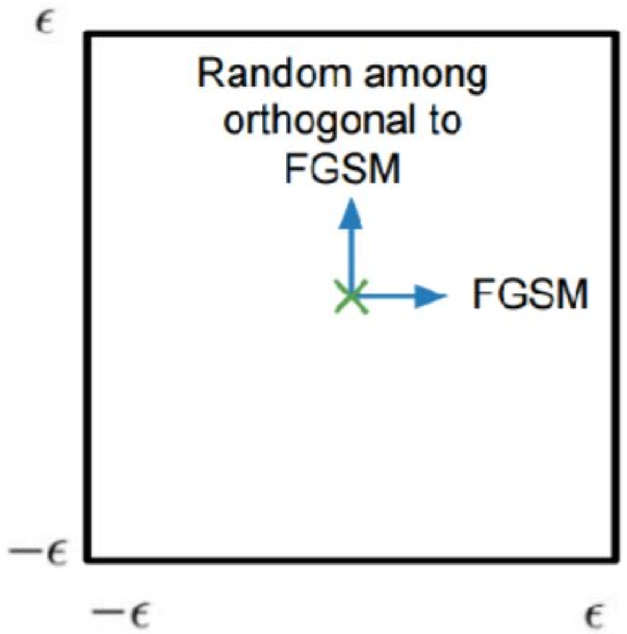
$$\|x\|_\infty = \max_i |x_i|$$

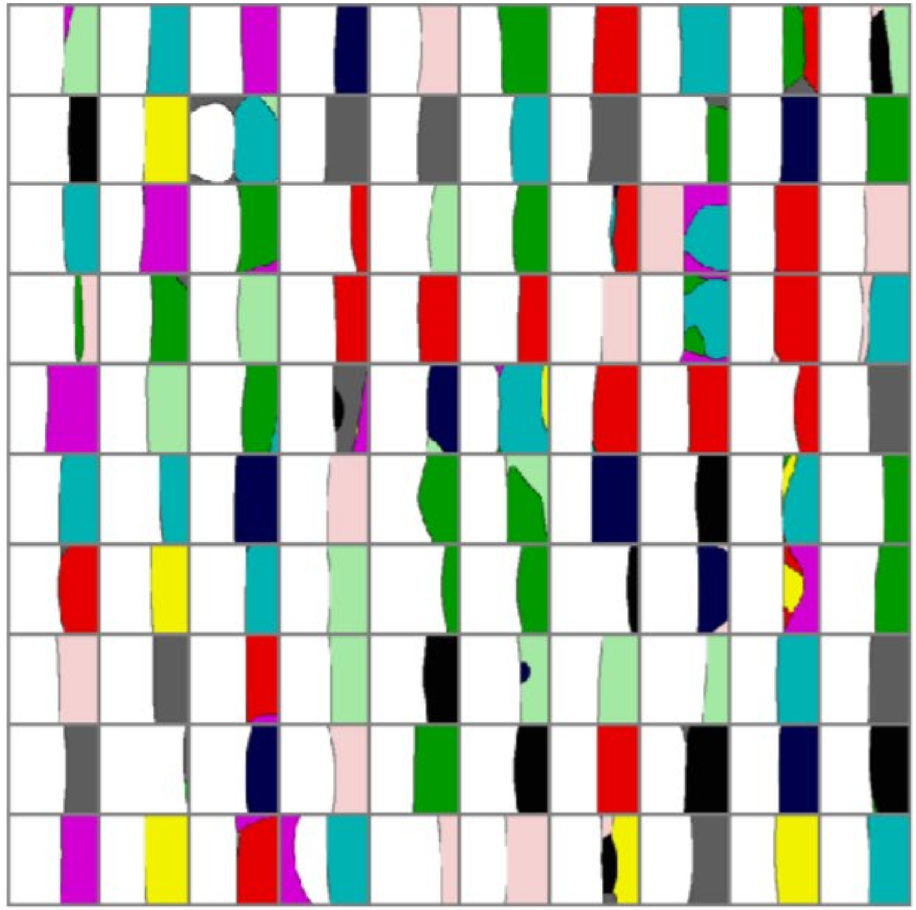
Infinity norm: maximal absolute value among all elements in x

$$x_{adv} = x + \varepsilon \cdot sign(\frac{\partial \ell(x;\boldsymbol{\theta})}{\partial x})$$

Explaining and Harnessing Adversarial Examples. 2015. https://arxiv.org/abs/1412.6572

Intuition from the fast gradient sign method



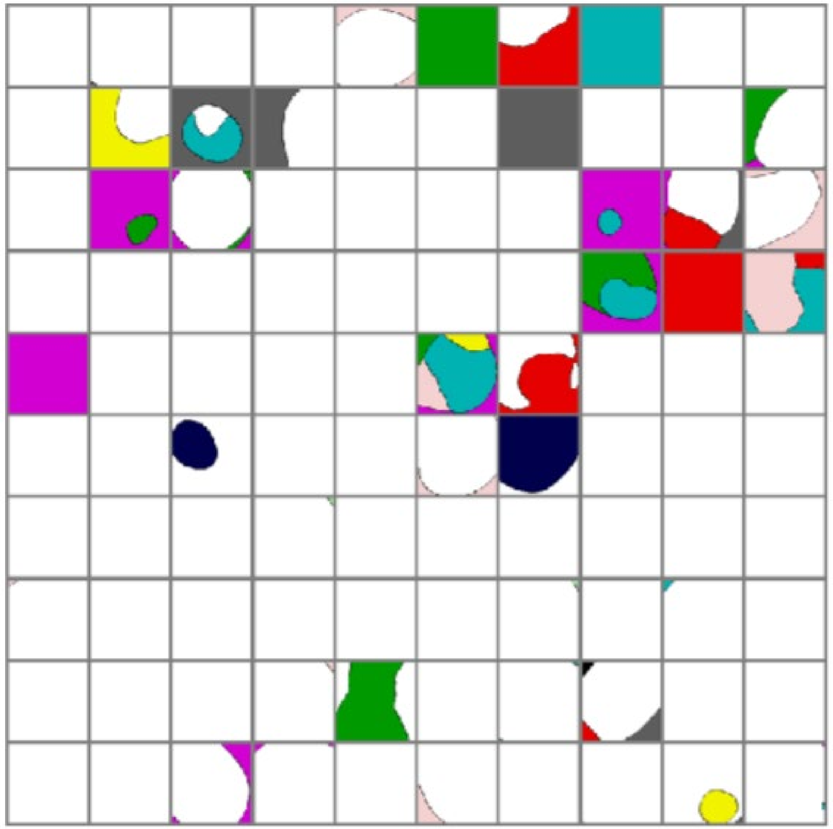


- Perturbation along the direction to increase loss fools the model effectively
- Adversarial samples are concentrated at least in some regions

Explaining and Harnessing Adversarial Examples. 2015. https://arxiv.org/abs/1412.6572

https://www.iangoodfellow.com/slides/2017-05-30-Stanford-cs213n.pdf

Intuition from the fast gradient sign method



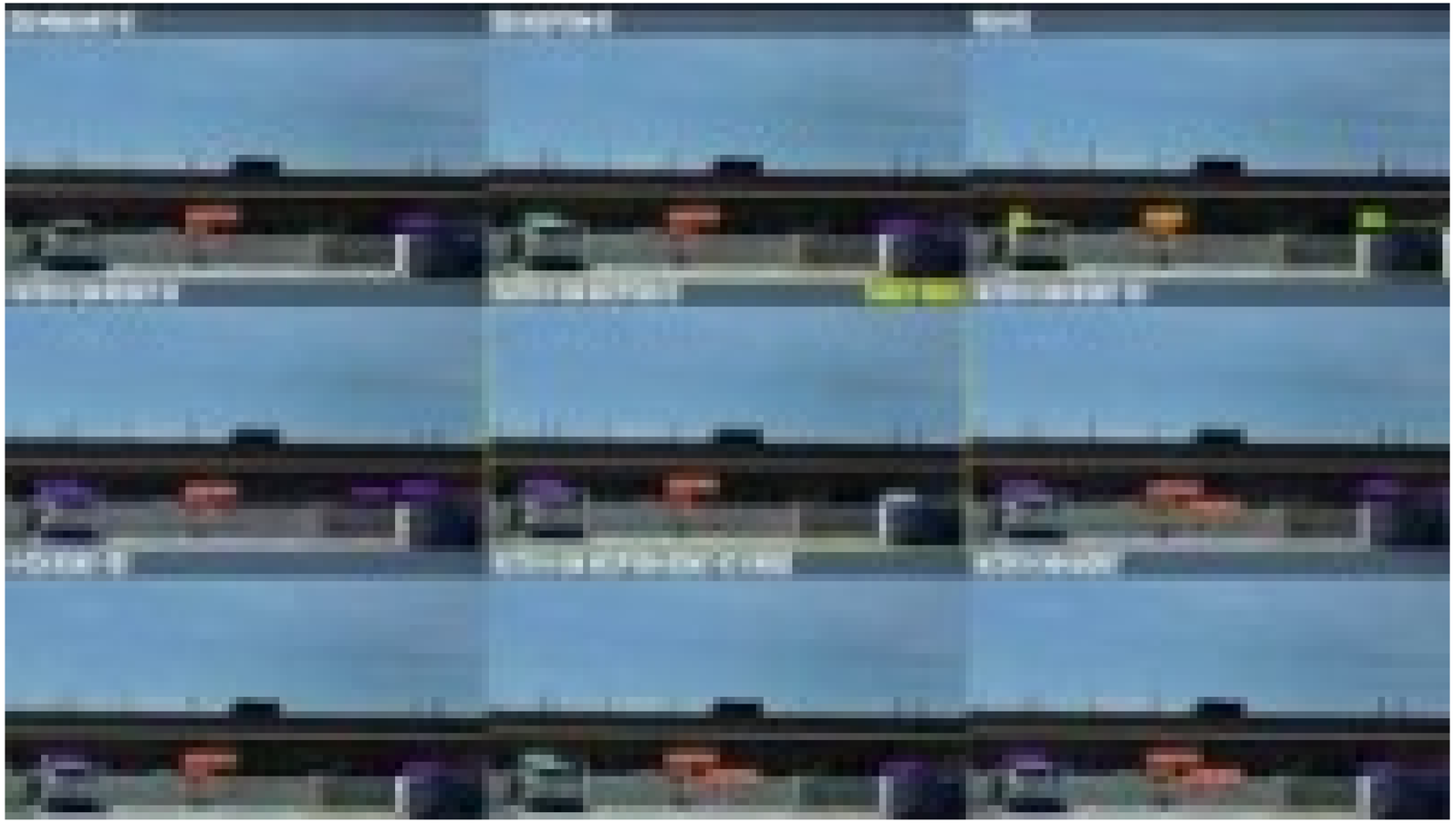
- Random perturbation is not effective to fool the model
- It indicates the adversarial samples are not randomly distributed, but rather on some continuous manifolds

Explaining and Harnessing Adversarial Examples. 2015. https://arxiv.org/abs/1412.6572

https://www.iangoodfellow.com/slides/2017-05-30-Stanford-cs213n.pdf

Adversarial examples with transferability

Adversarial samples created from one model can often fool other models. In other words, these attacking samples can be transferred between models

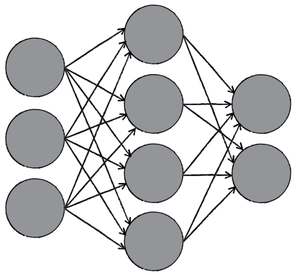


- Untargeted attack causing the missed detection to stop sign
- Test on 9 models
- Quite some models missed the stop sign, even the adversarial sample was prepared for a particular model

https://youtu.be/4KFhULX3v58

ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. 2019. https://arxiv.org/abs/1804.05810

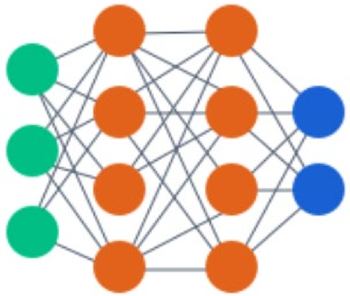
Adversarial black box attacking



Model working on domain A

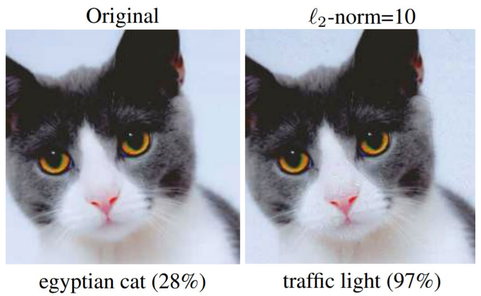
Hidden architecture and parameters

Collect data from the same or close domains



Train a model with common architecture

Create adversarial examples

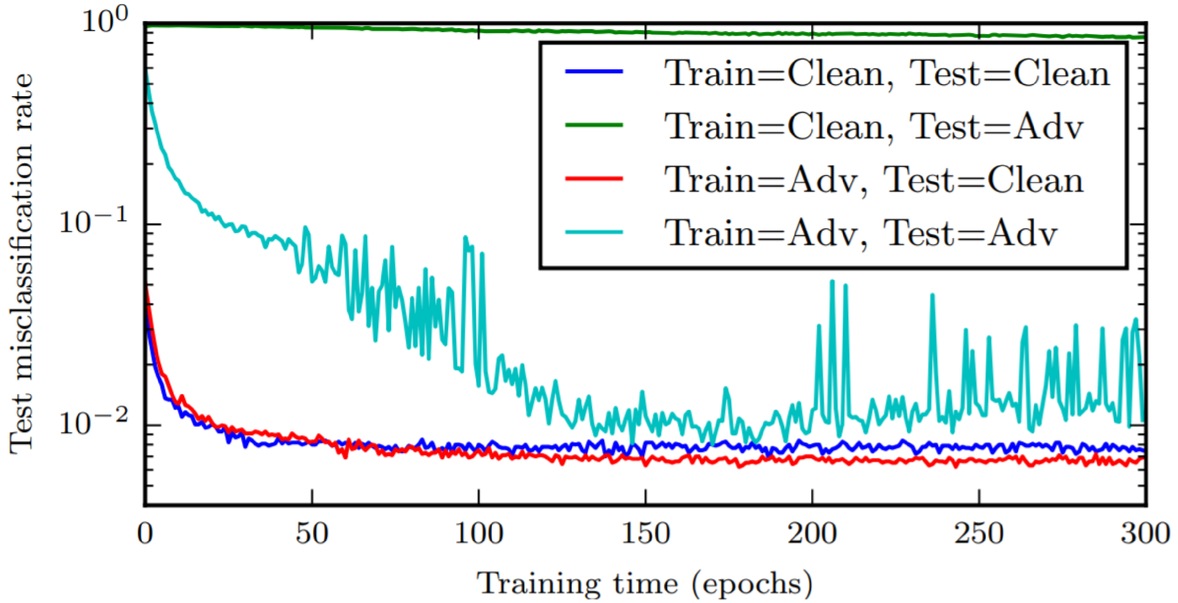


Attack the target model

Adversarial training

- Train the model with normal sample
- Prepare the adversarial samples using iterative or fast gradient sign methods
- Further train the model with normal and adversarial samples

$$\ell(x, x_{adv}; \hat{y}, \boldsymbol{\theta}) = \ell(x; \hat{y}, \boldsymbol{\theta}) + \ell(x_{adv}; \hat{y}, \boldsymbol{\theta})$$



Explaining and Harnessing Adversarial Examples. 2015. https://arxiv.org/abs/1412.6572 https://www.iangoodfellow.com/slides/2017-05-30-Stanford-cs213n.pdf

Attacking methods advance

Cup(16.48%)
Soup Bowl(16.74%)

Bassinet(16.59%)
Paper Towel(16.21%)

Teapot(24.99%)
Joystick(37.39%)

Hamster(35.79%)
Nipple(42.36%)

Original image (dog)

Airplane	Automobile	Bird
Cat	Deer	Frog
Horse	Ship	Truck

Target classes

- One-pixel attack
- Target and un-target perturbation are possible
- 16.04% of the ImageNet (ILSVRC 2012) test images can be perturbed to at least one target class by modifying just one pixel

One pixel attack for fooling deep neural networks. 2019. https://arxiv.org/abs/1710.08864

Outline

- Adversarial examples and adversarial training
- **Visualization of deep neural networks**

Motivation

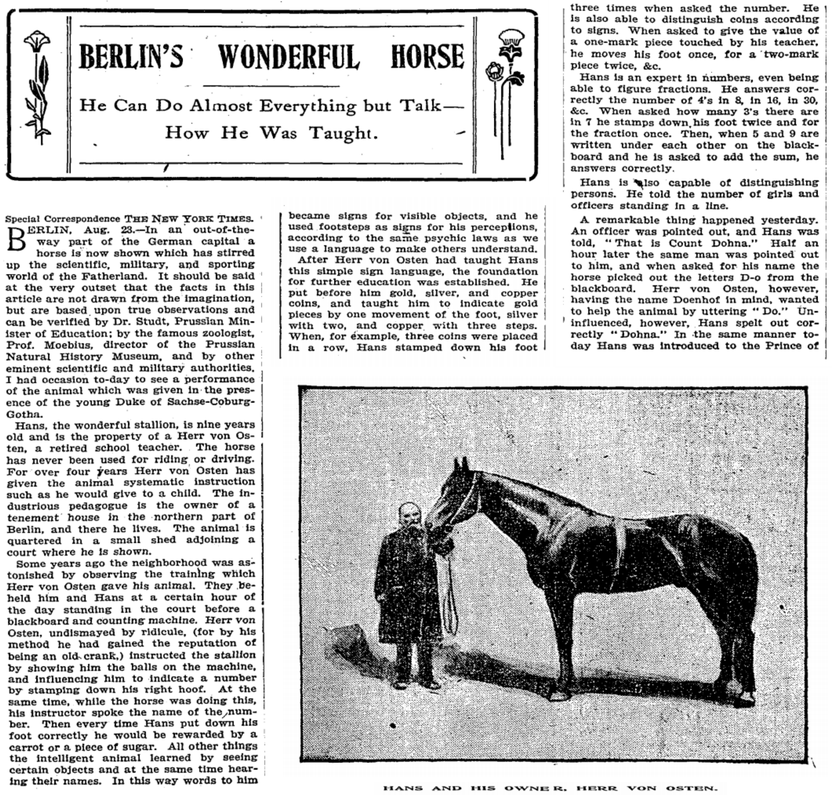
https://www.explainxkcd.com/wiki/index.php/1838:_Machine_Learning

- Curiosity
- Block-box solution may not be satisfying
- Difficulties in R&D
- Path to more general AI

“Machine learning' algorithms that can be reasonably described as pouring data into linear algebra and stirring until the output looks right” – from https://www.explainxkcd.com/wiki/index.php/1838:_Machine_Learning

Motivation

- problem solving does not mean intelligence in general



BERLIN'S WONDERFUL HORSE

He Can Do Almost Everything but Talk—How He Was Taught.

Special Correspondence THE NEW YORK TIMES.

BERLIN, Aug. 23.—In an out-of-the-way part of the German capital a horse is now shown which has stirred up the scientific, military, and sporting world of the Fatherland. It should be said at the very outset that the facts in this article are not drawn from the imagination, but are based upon true observations and can be verified by Dr. Studt, Prussian Minister of Education; by the famous zoologist, Prof. Moebius, director of the Prussian Natural History Museum, and by other eminent scientific and military authorities. I had occasion to-day to see a performance of the animal which was given in the presence of the young Duke of Sachse-Coburg-Gotha.

Hans, the wonderful stallion, is nine years old and is the property of a Herr von Osten, a retired school teacher. The horse has never been used for riding or driving. For over four years Herr von Osten has given the animal systematic instruction such as he would give to a child. The industrious pedagogue is the owner of a tenement house in the northern part of Berlin, and there he lives. The animal is quartered in a small shed adjoining a court where he is shown.

Some years ago the neighborhood was astonished by observing the training which Herr von Osten gave his animal. They beheld him and Hans at a certain hour of the day standing in the court before a blackboard and counting machine. Herr von Osten, undismayed by ridicule, (for by his method he had gained the reputation of being an old crank,) instructed the stallion by showing him the balls on the machine, and influencing him to indicate a number by stamping down his right hoof. At the same time, while the horse was doing this, his instructor spoke the name of the number. Then every time Hans put down his foot correctly he would be rewarded by a carrot or a piece of sugar. All other things the intelligent animal learned by seeing certain objects and at the same time hearing their names. In this way words to him became signs for visible objects, and he used footsteps as signs for his perceptions, according to the same psychic laws as we use a language to make others understand.

After Herr von Osten had taught Hans this simple sign language, the foundation for further education was established. He put before him gold, silver, and copper coins, and taught him to indicate gold pieces by one movement of the foot, silver with two, and copper with three steps. When, for example, three coins were placed in a row, Hans stamped down his foot three times when asked the number. He is also able to distinguish coins according to signs. When asked to give the value of a one-mark piece touched by his teacher, he moves his foot once, for a two-mark piece twice, &c.

Hans is an expert in numbers, even being able to figure fractions. He answers correctly the number of 4's in 8, in 16, in 30, &c. When asked how many 3's there are in 7 he stamps down his foot twice and for the fraction once. Then, when 5 and 9 are written under each other on the blackboard and he is asked to add the sum, he answers correctly.

Hans is also capable of distinguishing persons. He told the number of girls and officers standing in a line.

A remarkable thing happened yesterday. An officer was pointed out, and Hans was told, "That is Count Dohna." Half an hour later the same man was pointed out to him, and when asked for his name the horse picked out the letters D-o from the blackboard. Herr von Osten, however, having the name Doenhof in mind, wanted to help the animal by uttering "Do." Uninfluenced, however, Hans spelt out correctly "Dohna." In the same manner today Hans was introduced to the Prince of

HANS AND HIS OWNER, HERR VON OSTEN.

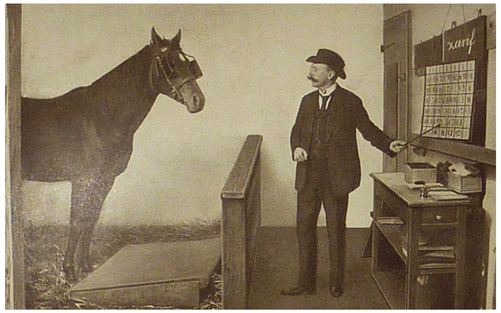
his foot down he snaps for the delicacy in the hand of his master. I doubt whether the horse really takes pleasure in his studies. He follows entirely mental impressions which he receives from the surroundings and which satisfy his wants."

Hans is the second horse Herr von Osten has trained. He claims that any horse of fair intelligence can be so taught. Herr von Osten's training is done purely from a scientific standpoint, and he told me that he greatly regretted the premature publicity given to his work. By the time this article is in print the Kaiser, who has heard with interest of this horse prodigy, will have seen the animal. EDWARD T. HEYN.

The New York Times
Published: September 4, 1904
Copyright © The New York Times

"BERLIN'S WONDERFUL HORSE; He Can Do Almost Everything but Talk—How He Was Taught" (PDF). *The New York Times*. 1904-09-04. Retrieved 2008-02-26.

https://en.wikipedia.org/wiki/Clever_Hans#cite_note-nytimes1904-3

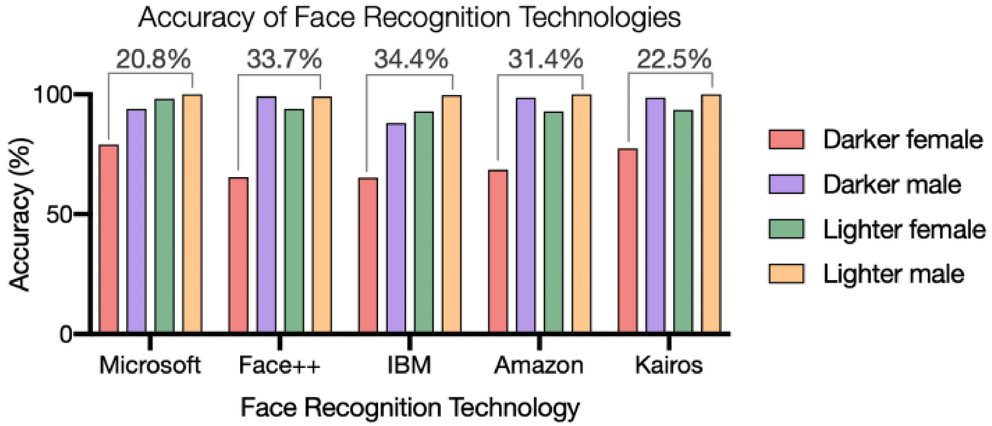


https://www.damninteresting.com/clever-hans-the-math-horse/

- Animal intelligence to solve math problem?
- Learn from surrounding human's face gestures
- Are NN the same horse?

Motivation

- Fairness and bias



Accuracy varies for gender and race

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification.
http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

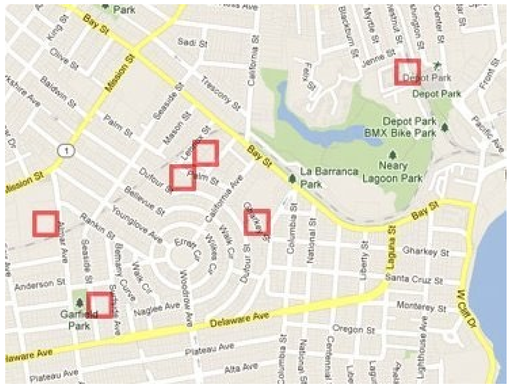
- Need to understand “black-box” to build fair and trustworthy AI system
- Detect and alarm data bias
- Control AI system within the range of consent

https://www.appsstore.ai/app-details/face-mask/?utm_source=facemaskpagelht

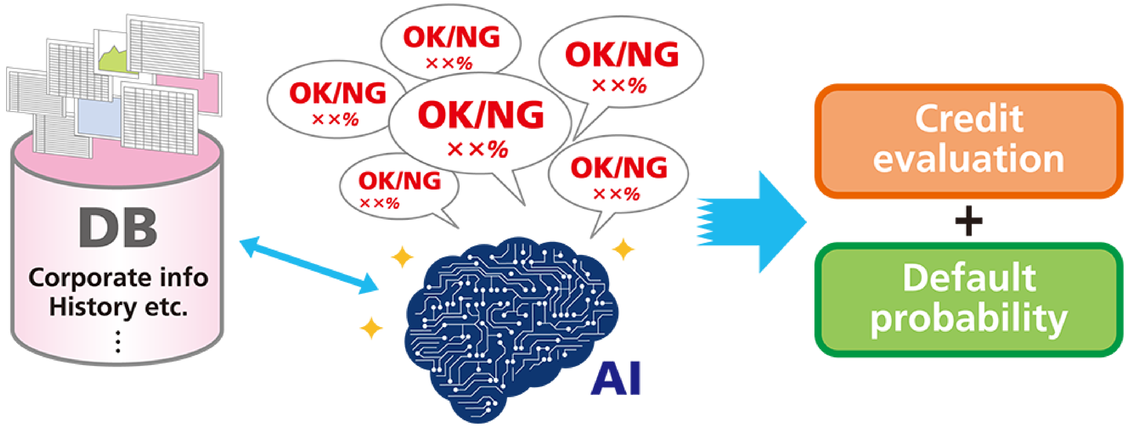
Do you agree?

Motivation

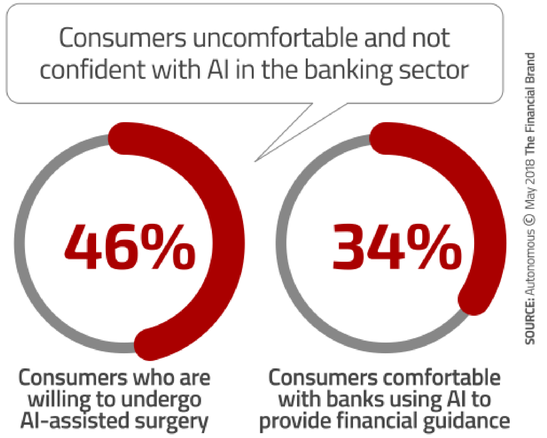
- Business requirement, legal requirement
- There are cases where decision/prediction made mainly by an AI system needs to explained to human



AI for Crime Prediction
http://www.predpol.com/



AI for loan underwriting



2018, https://thefinancialbrand.com/

Motivation

- Research needs

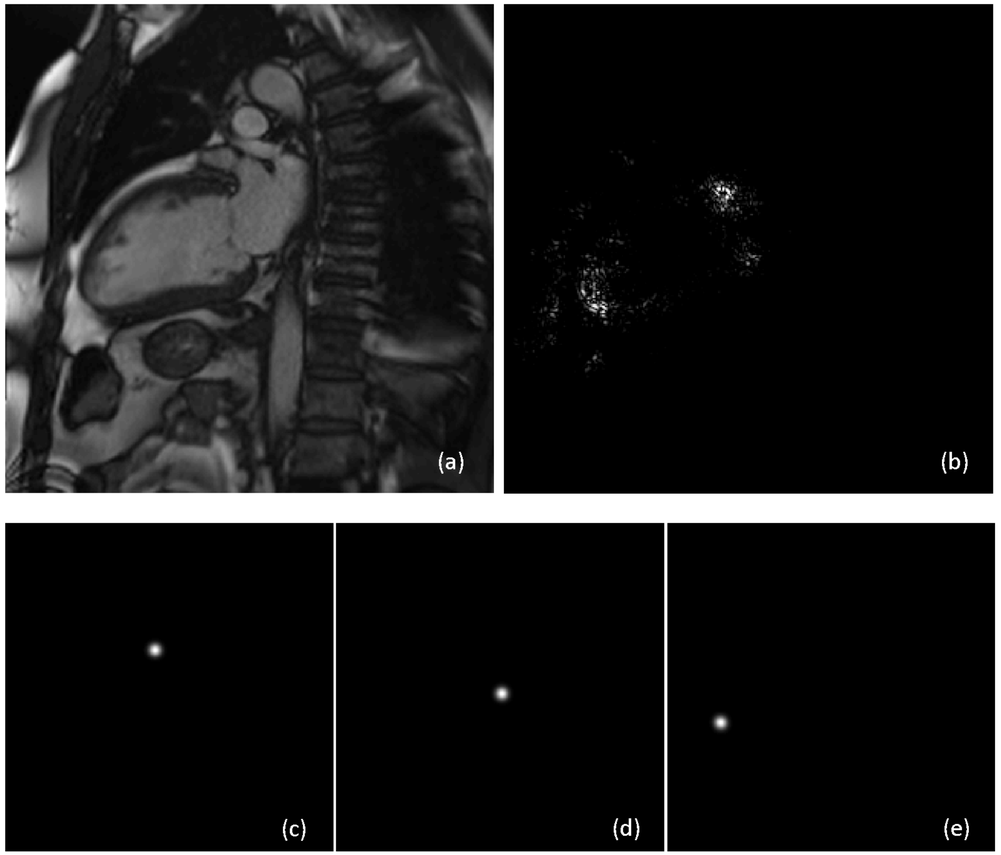
2. The authors also mentioned “transfer learning” from cine MRI, but the details are lacking. As described in the manuscript, there are cine-specific, LGE-specific, and T1-specific landmark detectors. The proposed work is, therefore not “generic” as claimed by the authors (page 5). From a technical point of view, there should be some more in-depth discussion on the type of features learned by the CNN, offering some explainability to readers of Radiology Artificial Intelligence, see Reyes et al. On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities. Radiology AI 2020

R1-2: Since more cine data was curated than LGE and T1 mapping, a transfer learning strategy was performed, where the neural networks were first trained with cine data as the pre-trained model. The LGE or T1 data was used to fine-tune the pre-trained model with reduced learning rate. The motivation is to utilize the bigger cine datasets to help other tasks. Manuscript was modified to give more information about the transfer learning strategy used in this study. The paper was further modified to remove the “generic” claims and discussed further about extension to other detection tasks.

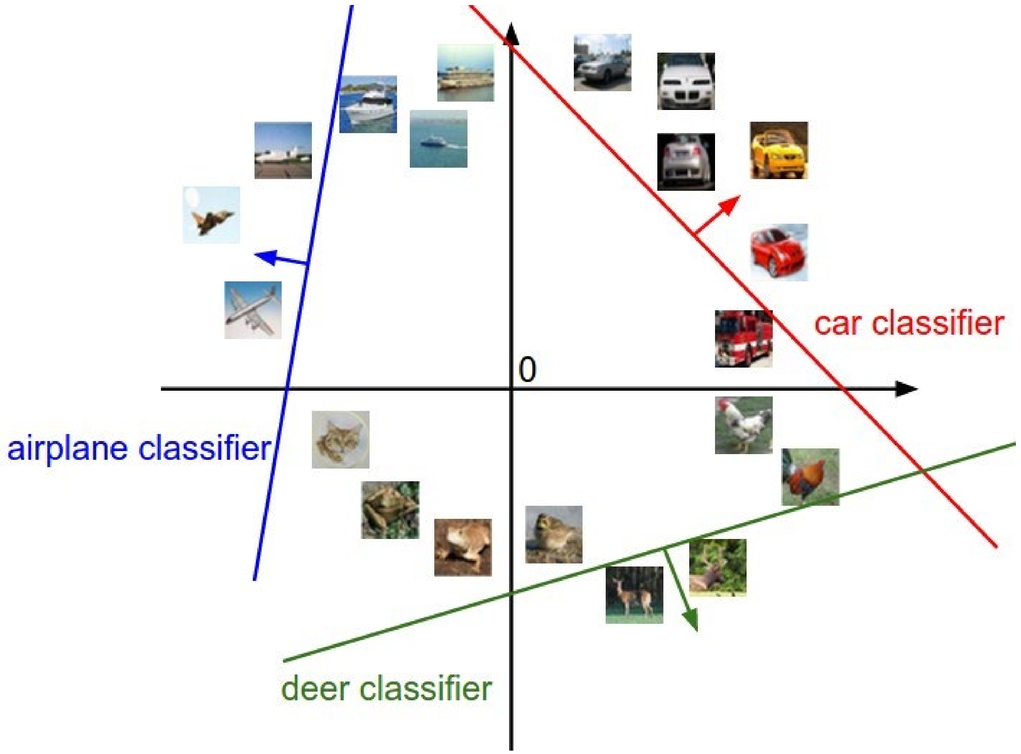
More discussion was added for explainability of this model. To demonstrate the features where model learned during training, the saliency maps were further computed for representative samples, added as an appendix section.

Example of a reviewer’s comment

Appendix Figure E6: Saliency maps for landmark detection.



Simple models are more interpretable

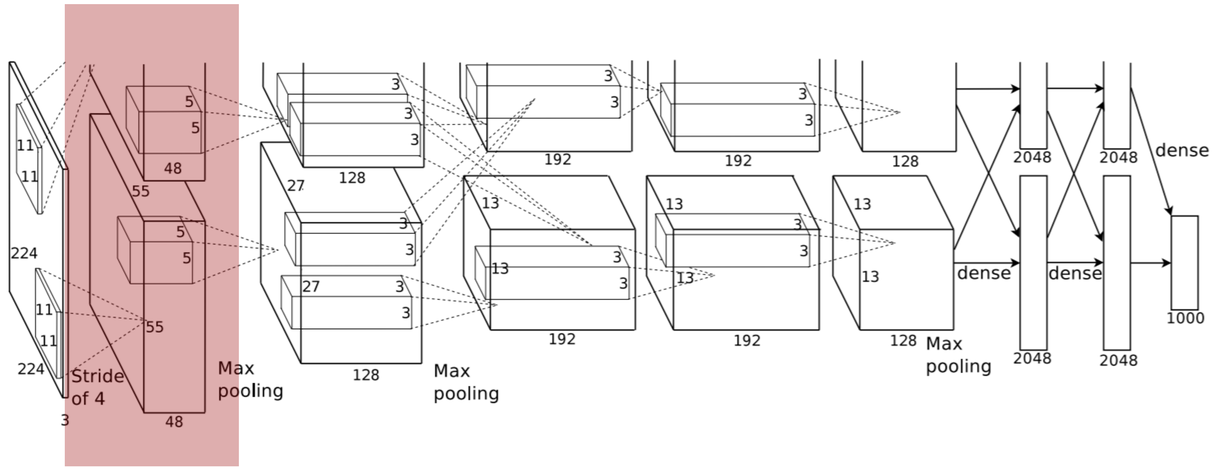


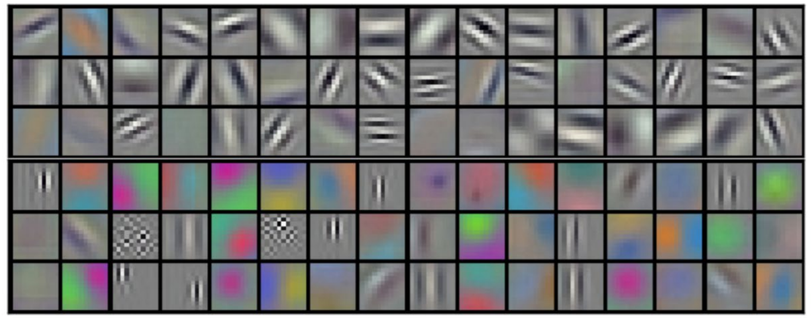
$$y = \sigma(\boldsymbol{W}\boldsymbol{X} + \boldsymbol{b})$$

- Decision boundary is linear – hyperplane
- Visualizing the weights gives clue for every class – template matching

Figures from https://cs231n.github.io/linear-classify/

First layer weights





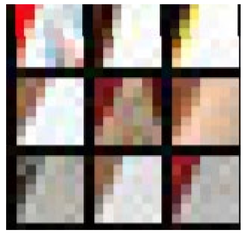
11x11, Input channel 3, output channel 64

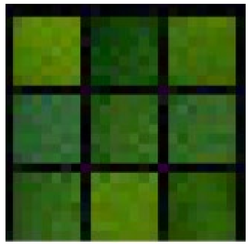
ImageNet Classification with Deep Convolutional Neural Networks. NIPS 2012
Deep Residual Learning for Image Recognition. CVPR 2016.

Activation and corresponding image patches

activation

Image patches triggering this activation



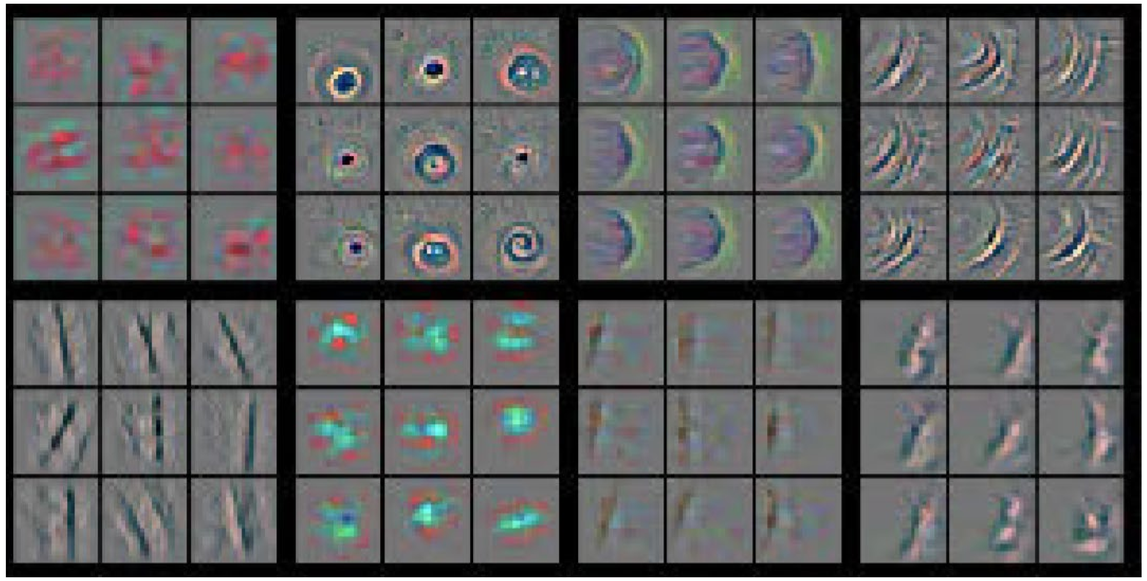


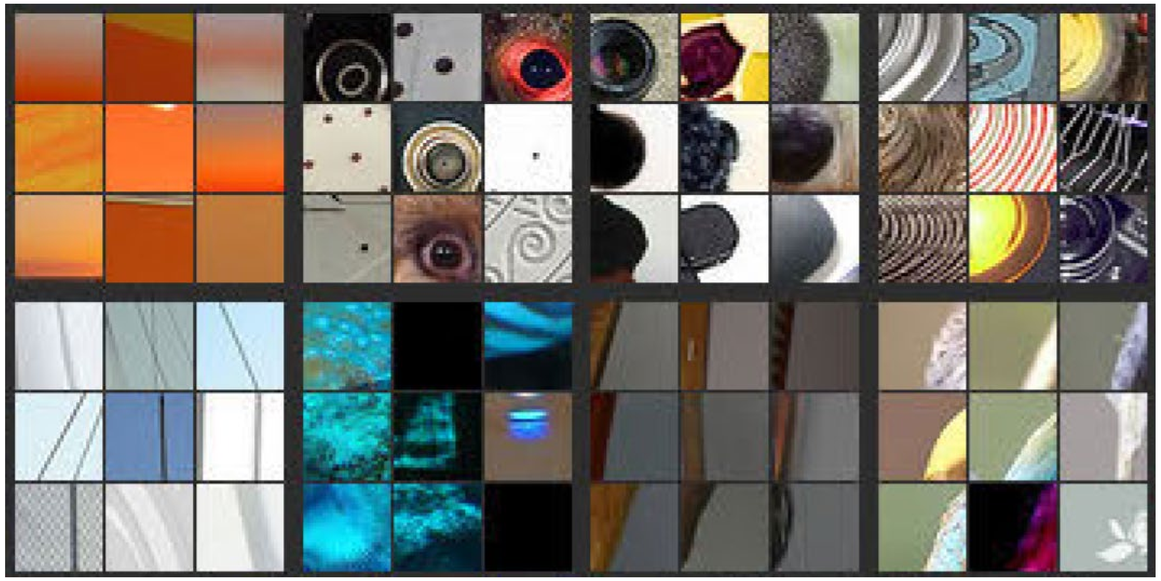
- Select a batch of image
- Pass images through the network
- Track the top activations and corresponding image patches (receptive fields)

First layer is simple image feature detector

Visualizing and Understanding Convolutional Networks. https://arxiv.org/abs/1311.2901

Deeper layers focus on higher level content

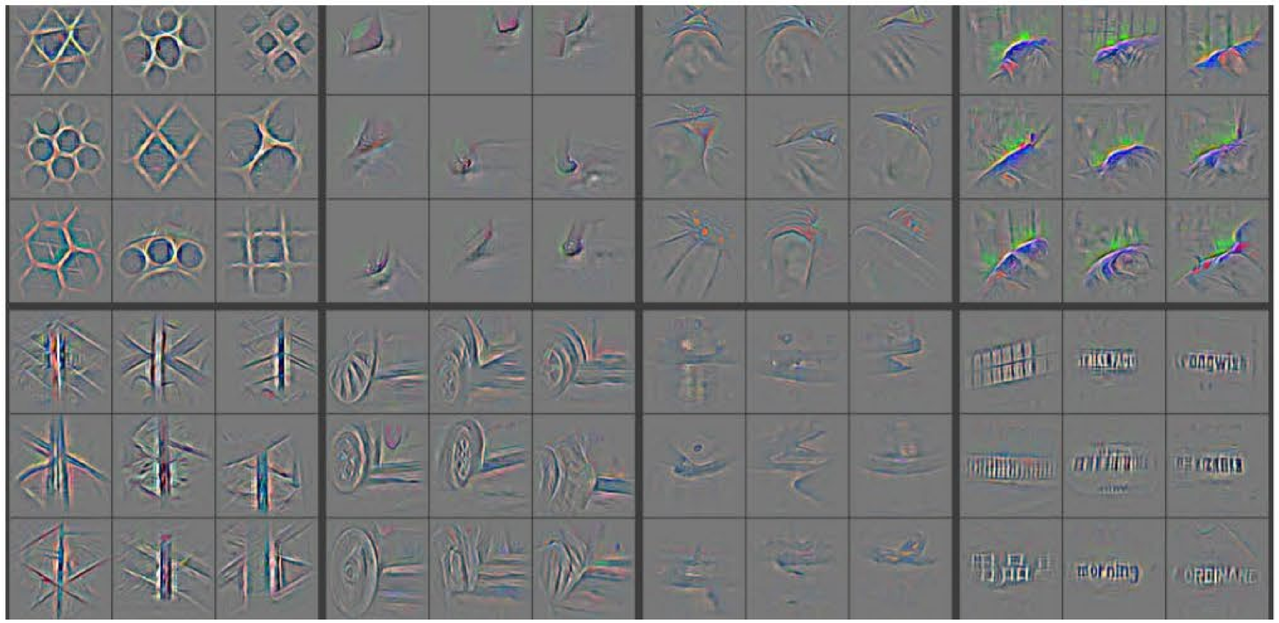


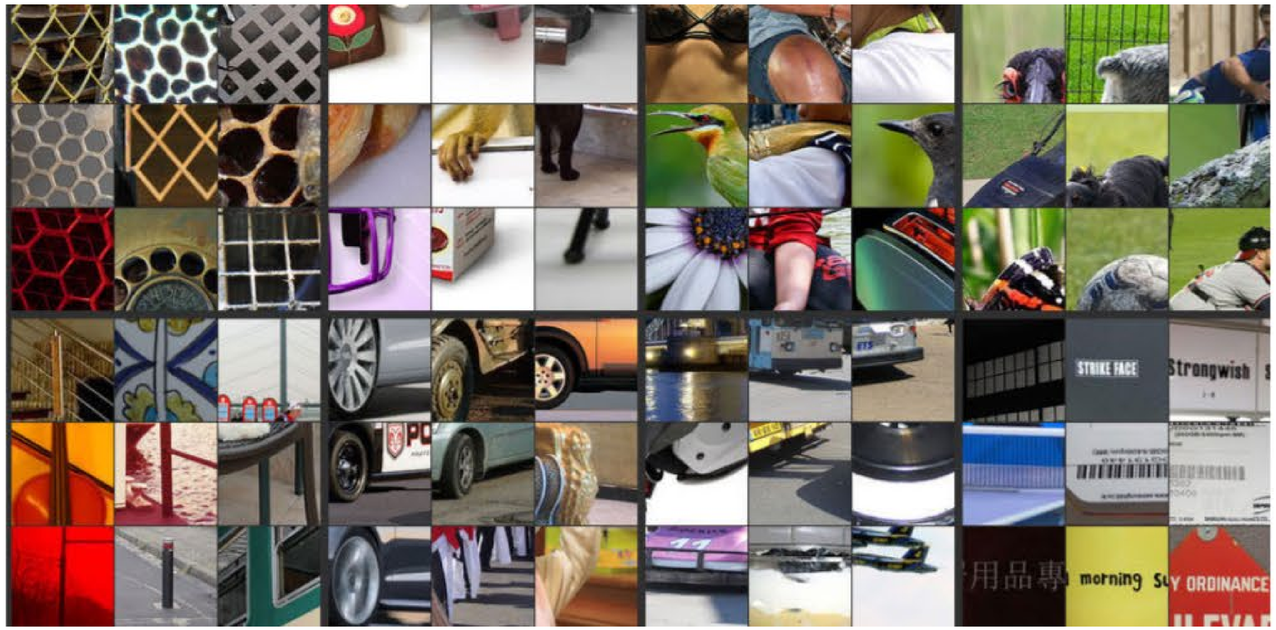


Intermediate layers are activated for compound image structures

Visualizing and Understanding Convolutional Networks. https://arxiv.org/abs/1311.2901

Deeper layers focus on higher level content

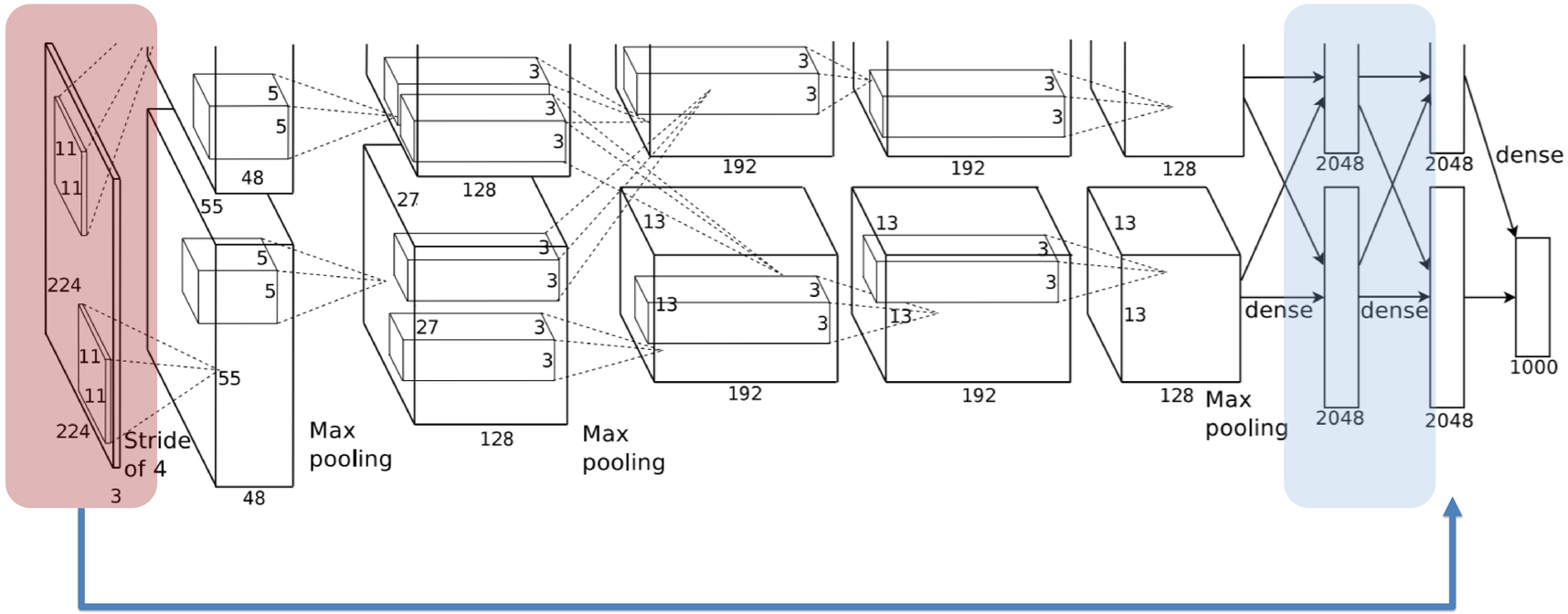




Later layers focus on global structures and image content

Visualizing and Understanding Convolutional Networks. https://arxiv.org/abs/1311.2901

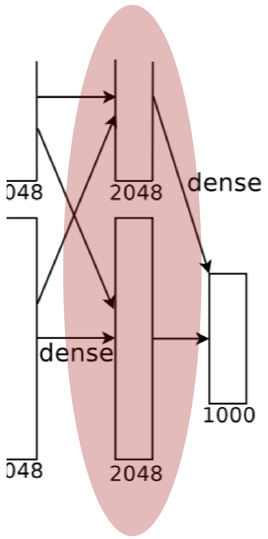
Output layer : compact representation



Neural network model compresses the image to a compact vector representation

e.g. Alex net, 224x224x3 -> 4096 ->1000

Output layer : nonlinear feature extraction

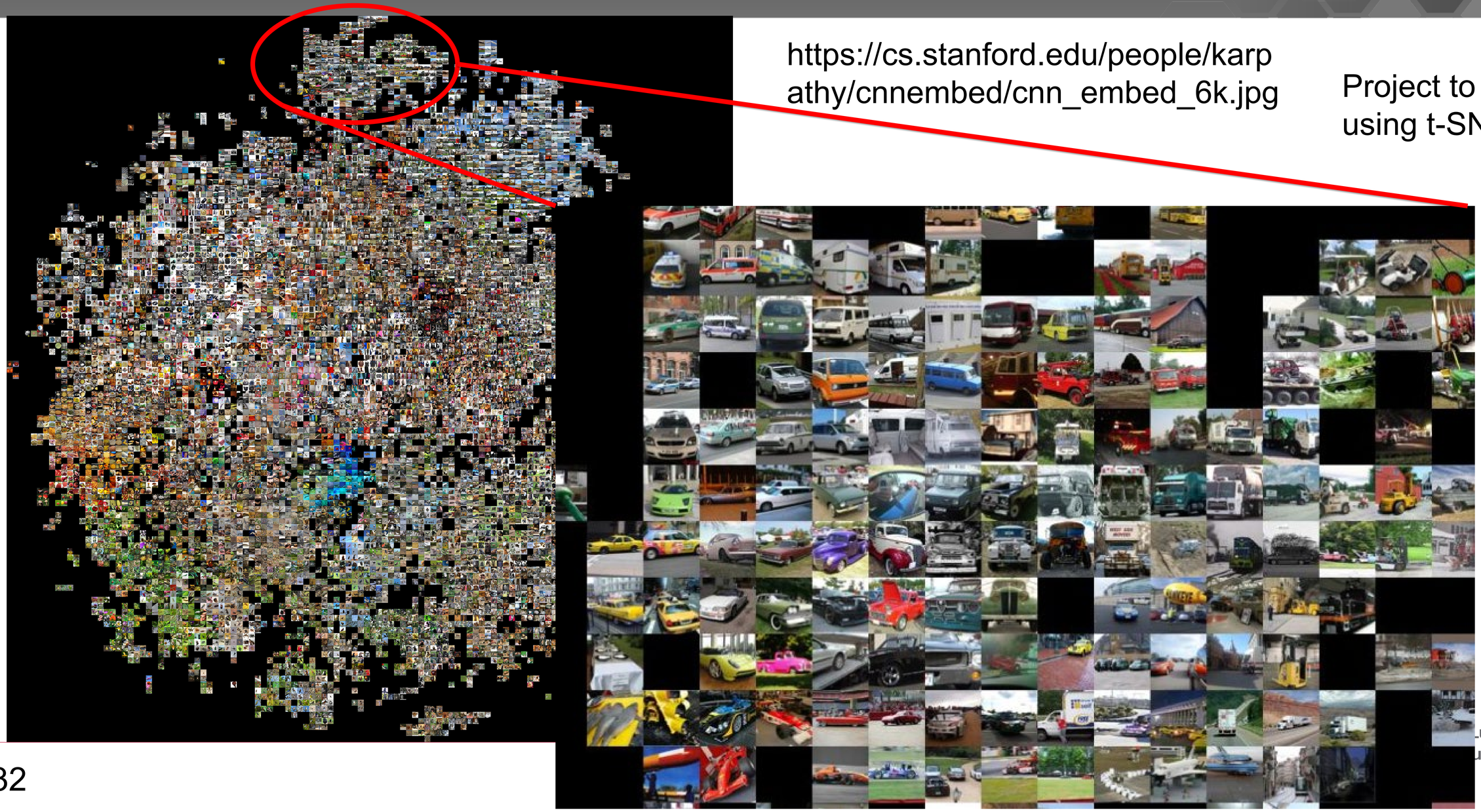


A linear classifier

- Neural network worked as a nonlinear feature extraction
- Convert complex image content to linearly separable feature vector!

Query images Images with closest feature vectors in L2 norm

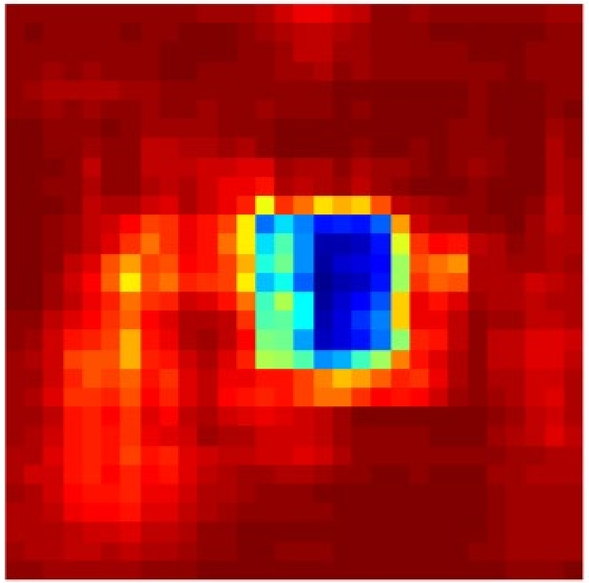
Output layer : nonlinear feature extraction

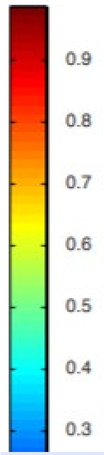


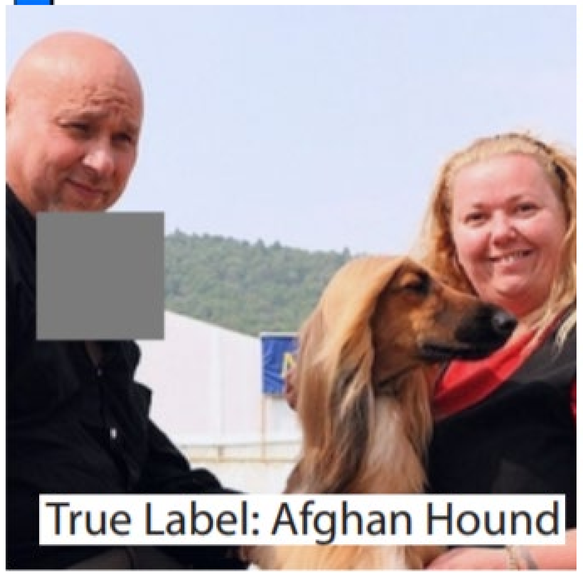
https://cs.stanford.edu/people/karpathy/cnnembed/cnn_embed_6k.jpg

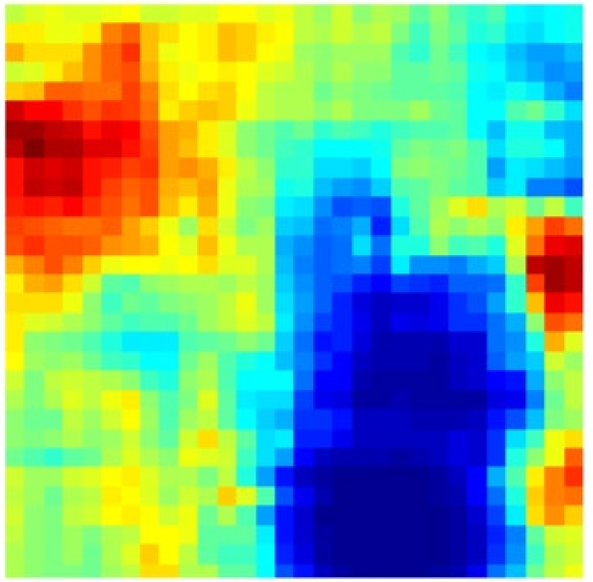
Project to 2D using t-SNE

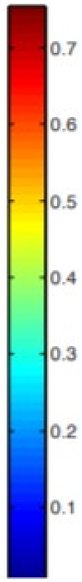
End-to-end visualization: Occlusion







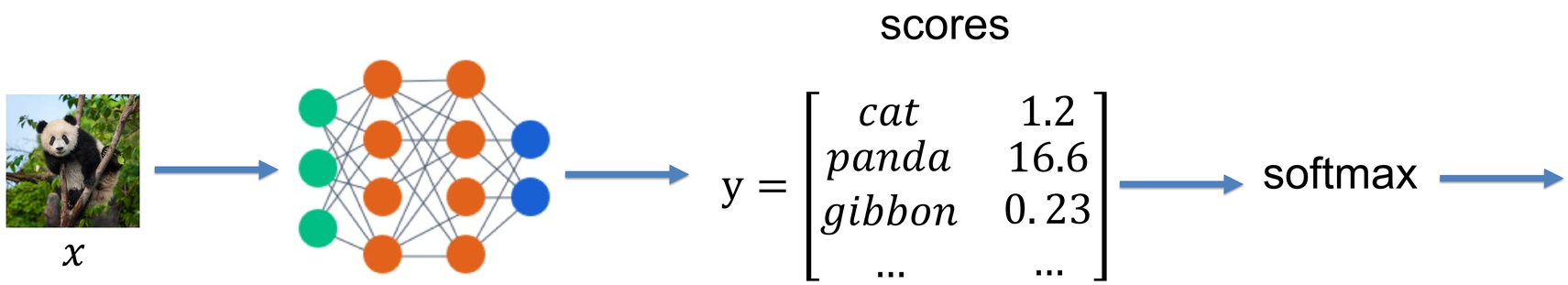




- Mask off some portion of image
- Compute prediction probability
- Sliding the mask across the image

Visualizing and Understanding Convolutional Networks. https://arxiv.org/abs/1311.2901

Saliency map for classification: Class Activation map



- More accurate the classifier, higher the corresponding score
- For every pixel in the image, how does it influence the score?

$$\left|\frac{\partial Score\_panda}{\partial x}\right|$$

For multi-channel inputs (e.g. rgb), take the maximal value across channels

Or, keep three channels still as r,g,b

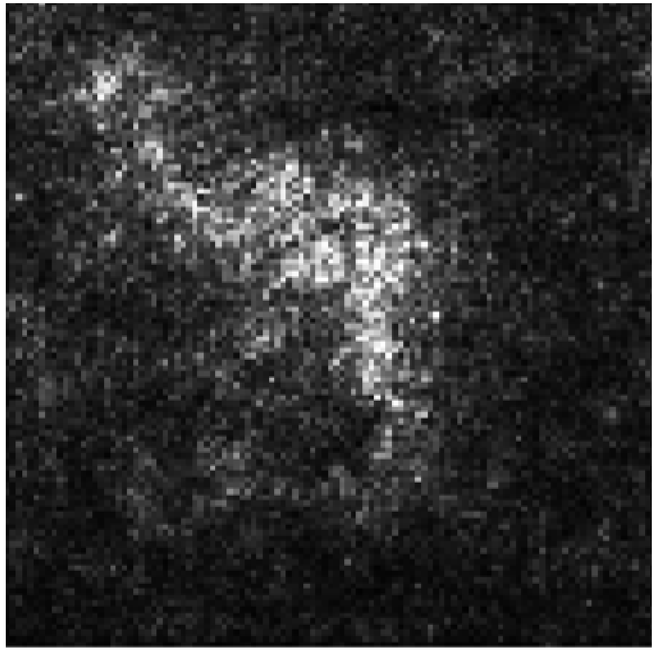
Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034v2

Saliency map for classification

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034v2

Saliency map with MonteCarlo

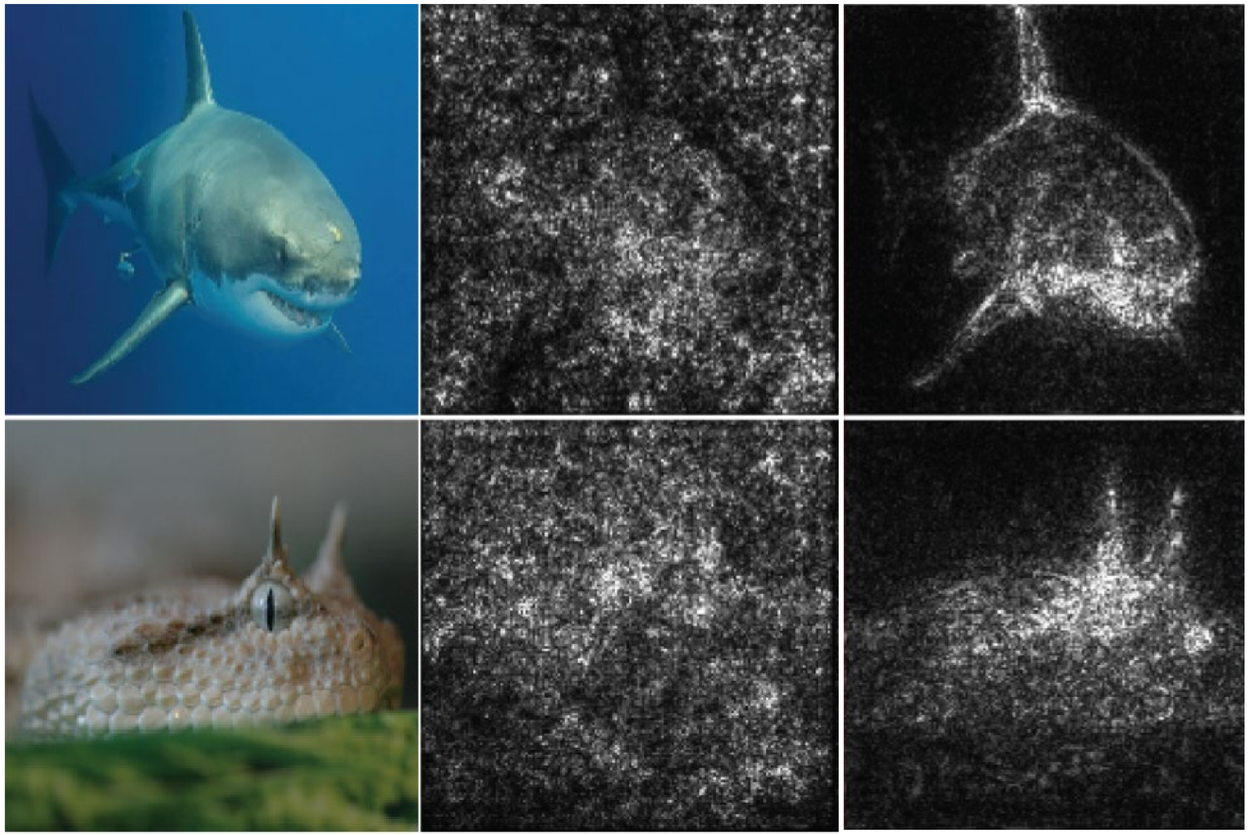
x



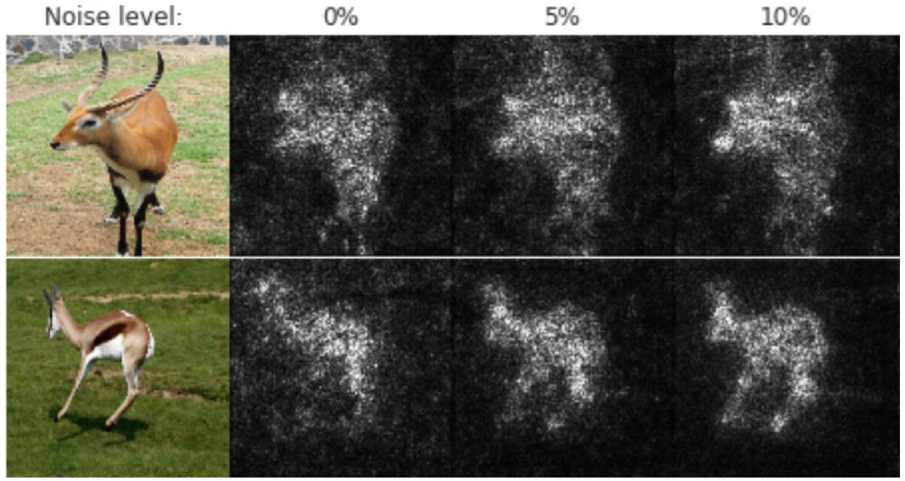
```
for i in range(50):
      saliency_map += compute(X + sigma*np.random.rand(X.shape))

saliency_map /= 50
```

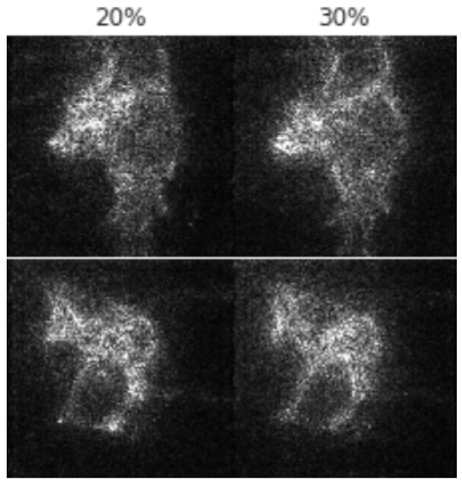
SmoothGrad: removing noise by adding noise. https://arxiv.org/abs/1706.03825

Saliency map with MonteCarlo



- Can be a useful trick

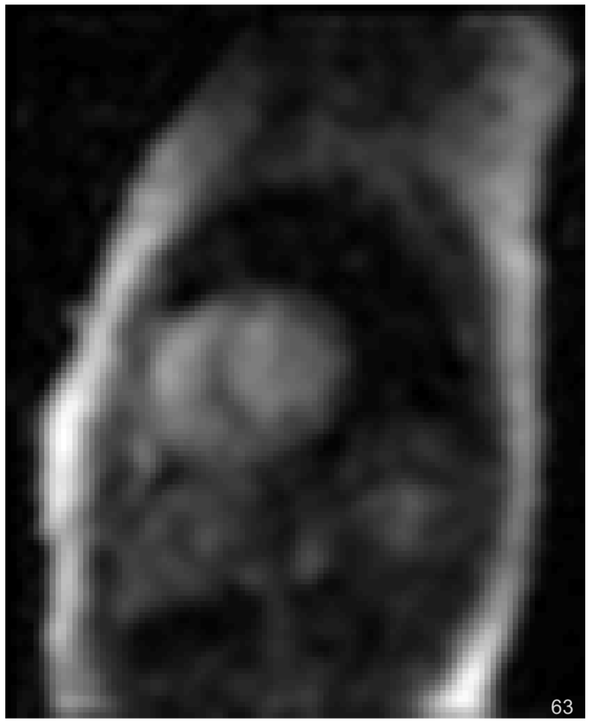


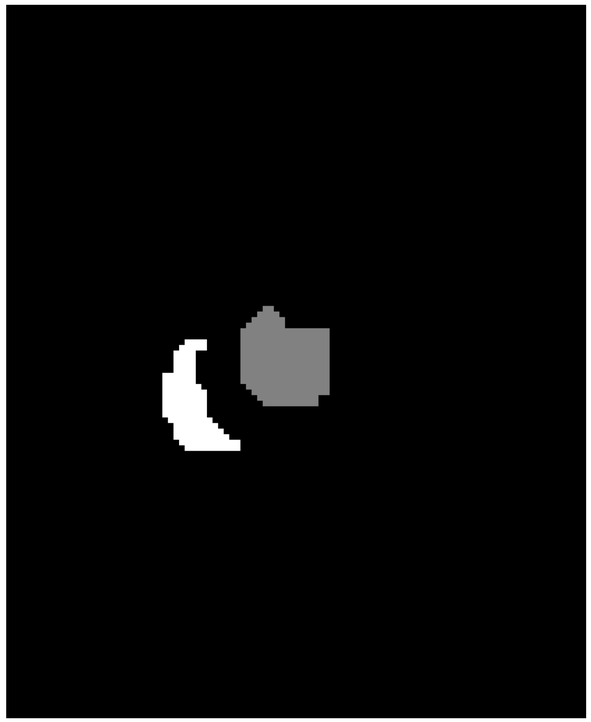


- May need to experiment different noise levels

SmoothGrad: removing noise by adding noise. https://arxiv.org/abs/1706.03825

Saliency map for segmentation





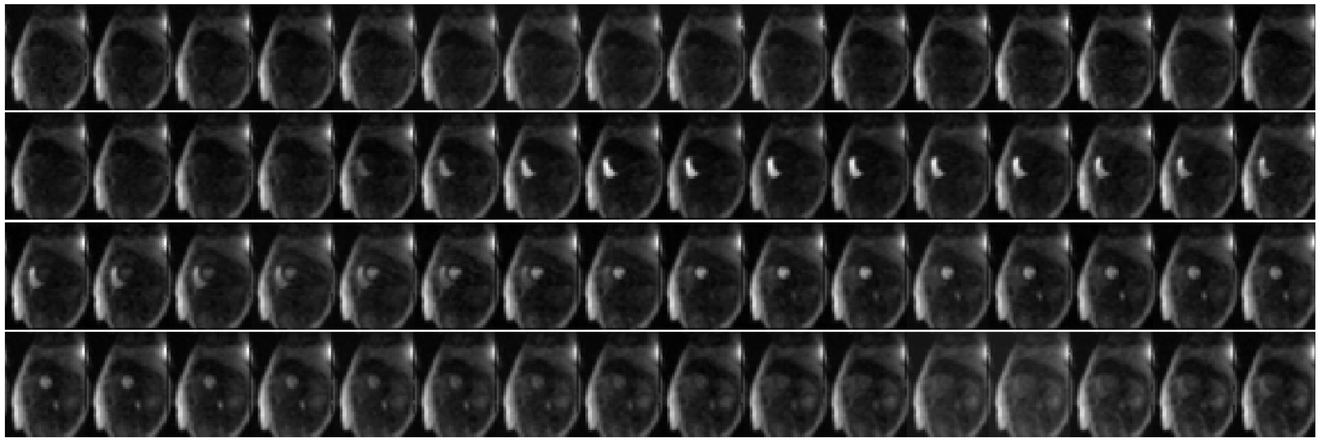
$$\ell_{seg}(x) = \frac{1}{N}\sum_{i=0}^{N-1} \ell_{ce}\left(x^{(i)}\right)$$

$$S(x) = \left|\frac{\partial \ell_{seg}}{\partial x^{(i)}}\right| \text{ for every pixel}$$

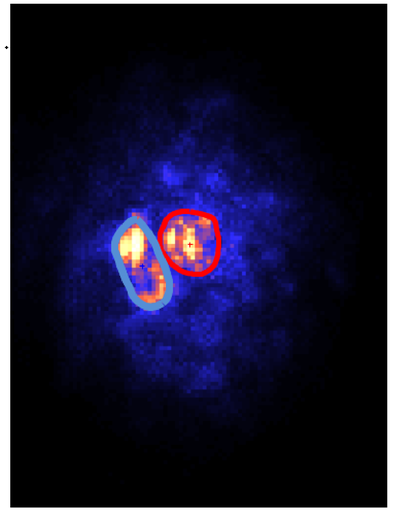
Compute with one backprop to input

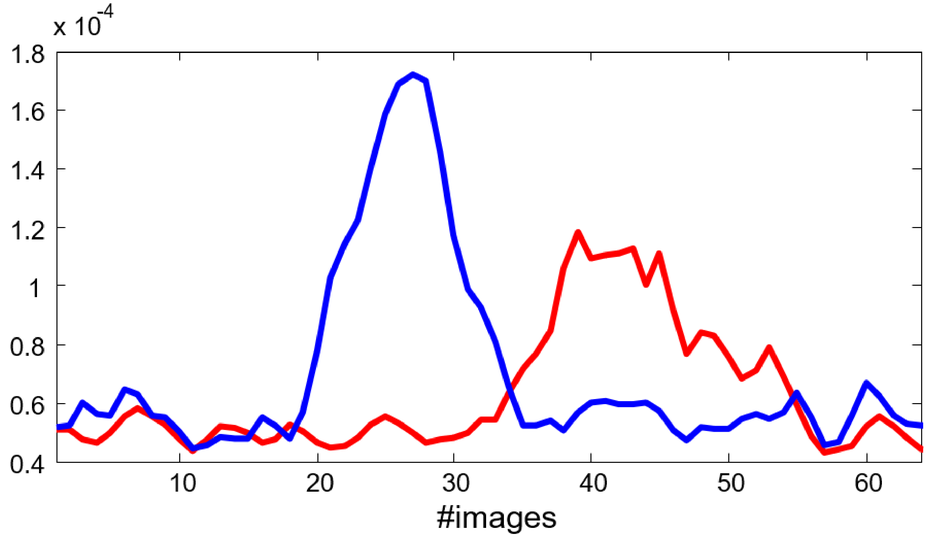
Automated detection of left ventricle in arterial input function images for inline perfusion mapping using deep learning: A study of 15,000 patients. 2020. https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.28291

Saliency map for segmentation



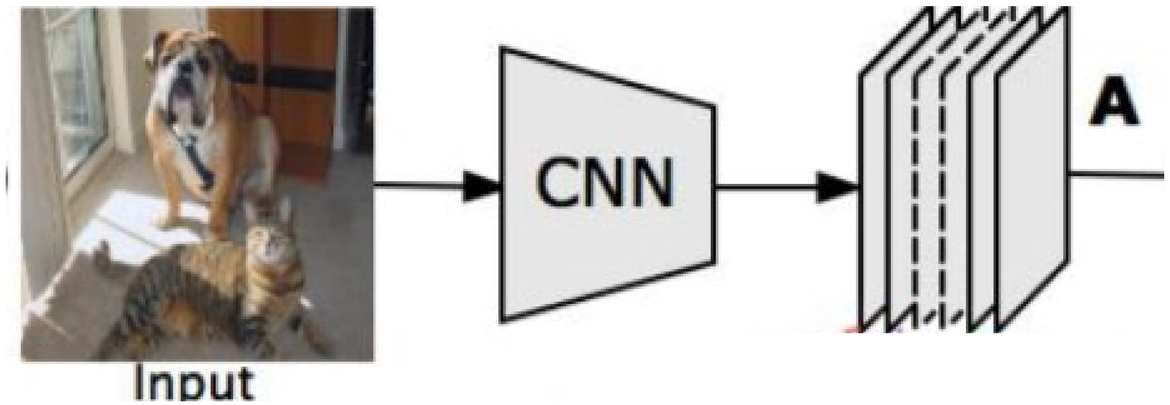
2D+T
Dynamic saliency map



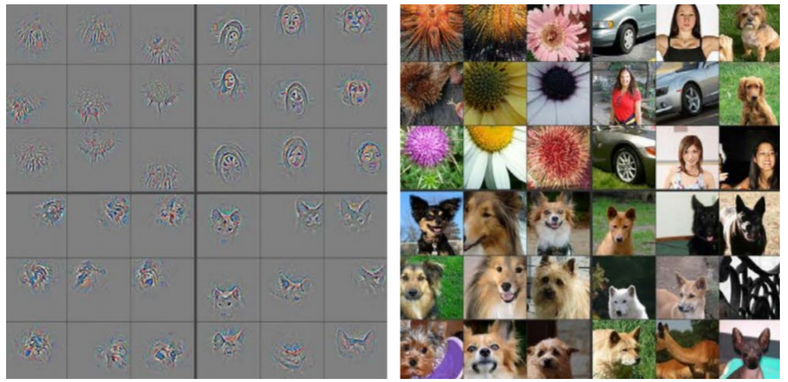


Automated detection of left ventricle in arterial input function images for inline perfusion mapping using deep learning: A study of 15,000 patients. 2020. https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.28291

Saliency map on the CONV layer: Grad-CAM



- CONV layer generate activations
- Focus on image content and global structure

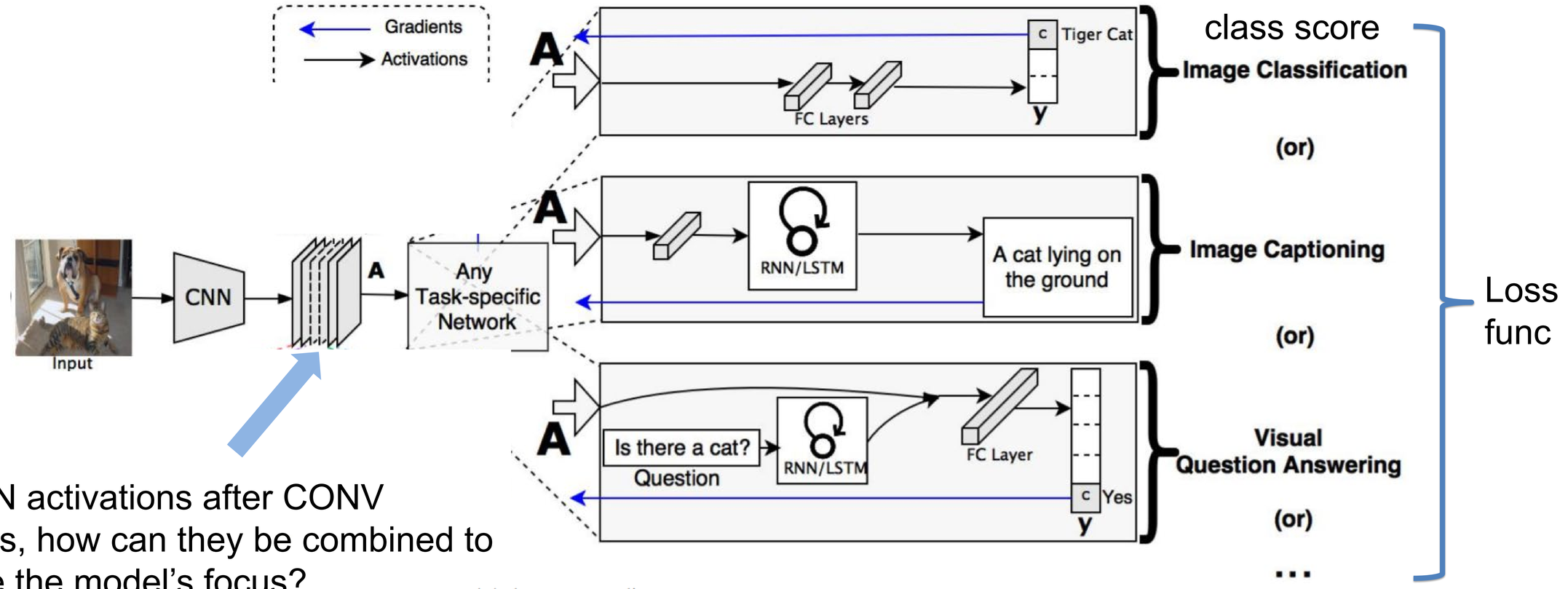


IDEA: activation after CONV modules tell where the model put its focus

Use CONV activation $A^k, k = 0, \ldots, N-1$, a total of N channels

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2019. https://arxiv.org/abs/1610.02391

Saliency map on the last CONV layer: Grad-CAM



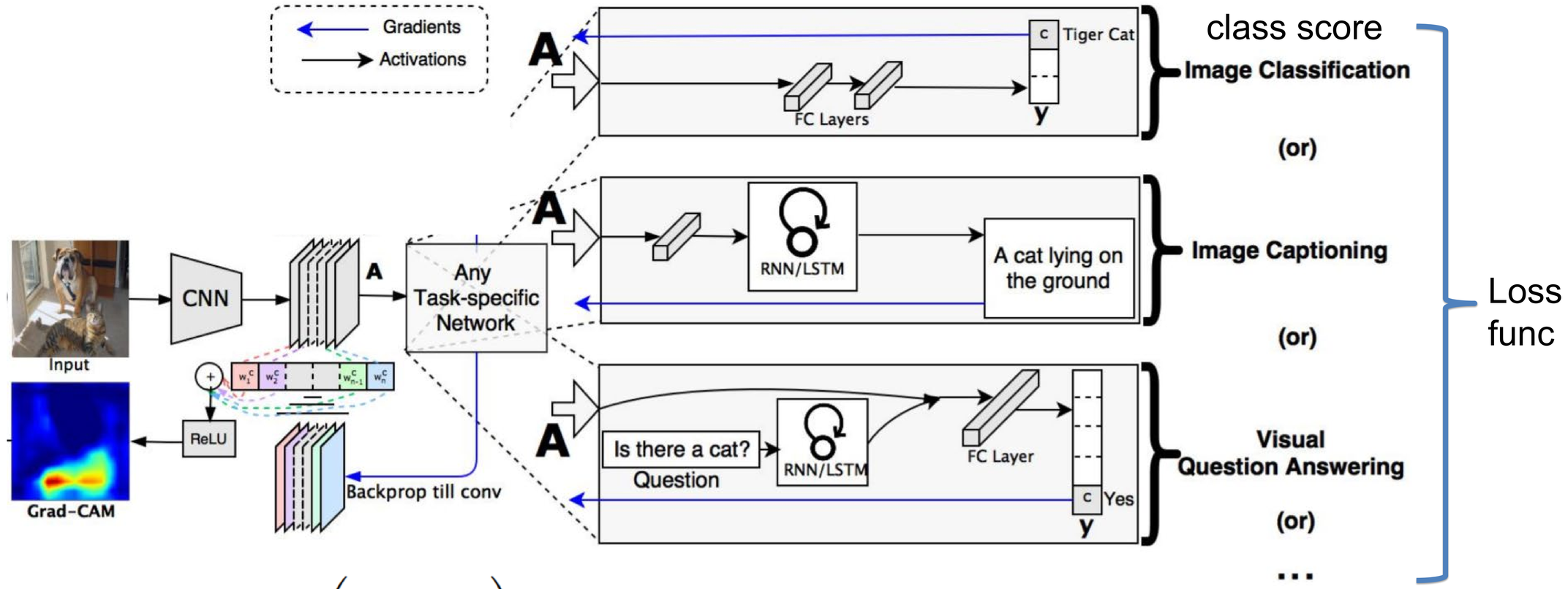
For all N activations after CONV modules, how can they be combined to indicate the model's focus?

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

- Compute weighing coefficient for every activation
- Use the derivative of class score or loss to A^k

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2019. https://arxiv.org/abs/1610.02391

Saliency map on the last CONV layer: Grad-CAM

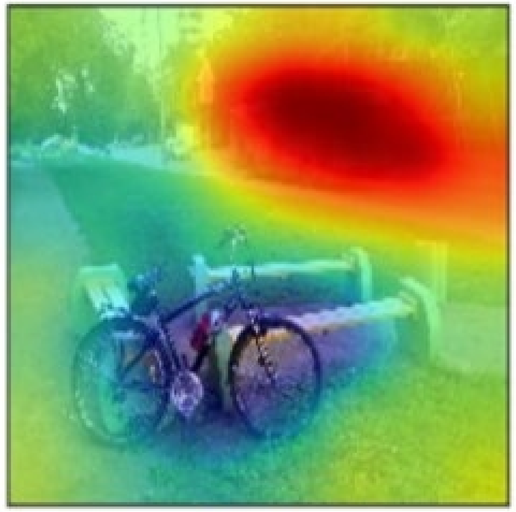


$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

- Weighted sum of activations
- Go through a ReLU, since we want score to be high positive number for correct class

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2019. https://arxiv.org/abs/1610.02391

Saliency map on the last CONV layer: Grad-CAM



bus

- Gain popularity, especially in imaging applications
- Flexible for many architectures using CNN
- Can create visualization for activation outputs in the middle of network
- Often require interpolation to upsample the visualization

Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. 2019. https://arxiv.org/abs/1610.02391

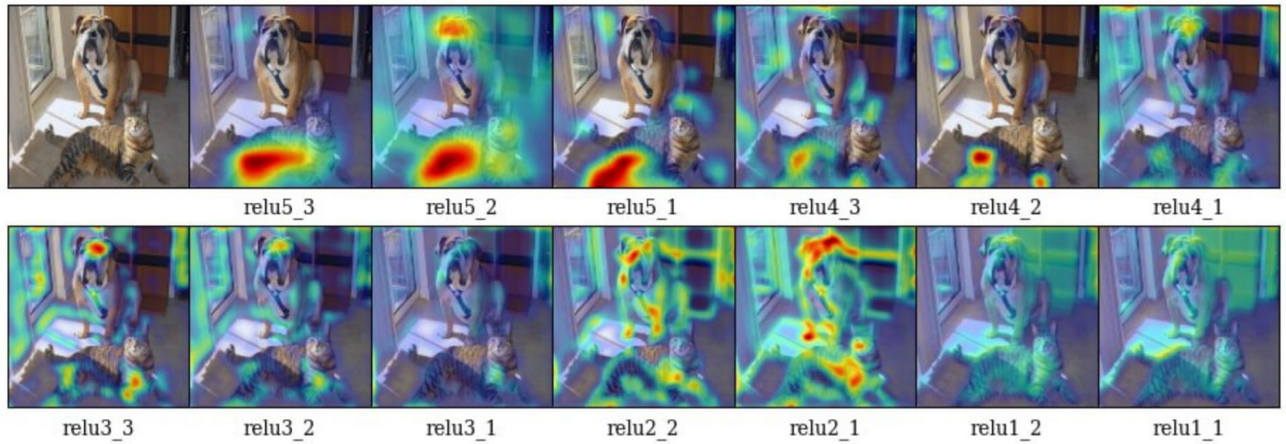


Fig. 13: Grad-CAM at different convolutional layers for the 'tiger cat' class. This figure analyzes how localizations change qualitatively as we perform Grad-CAM with respect to different feature maps in a CNN (VGG16 [52]). We find that the best looking visualizations are often obtained after the deepest convolutional layer in the network, and localizations get progressively worse at shallower layers. This is consistent with our intuition described in Section 3 of main paper, that deeper convolutional layer capture more semantic concepts.

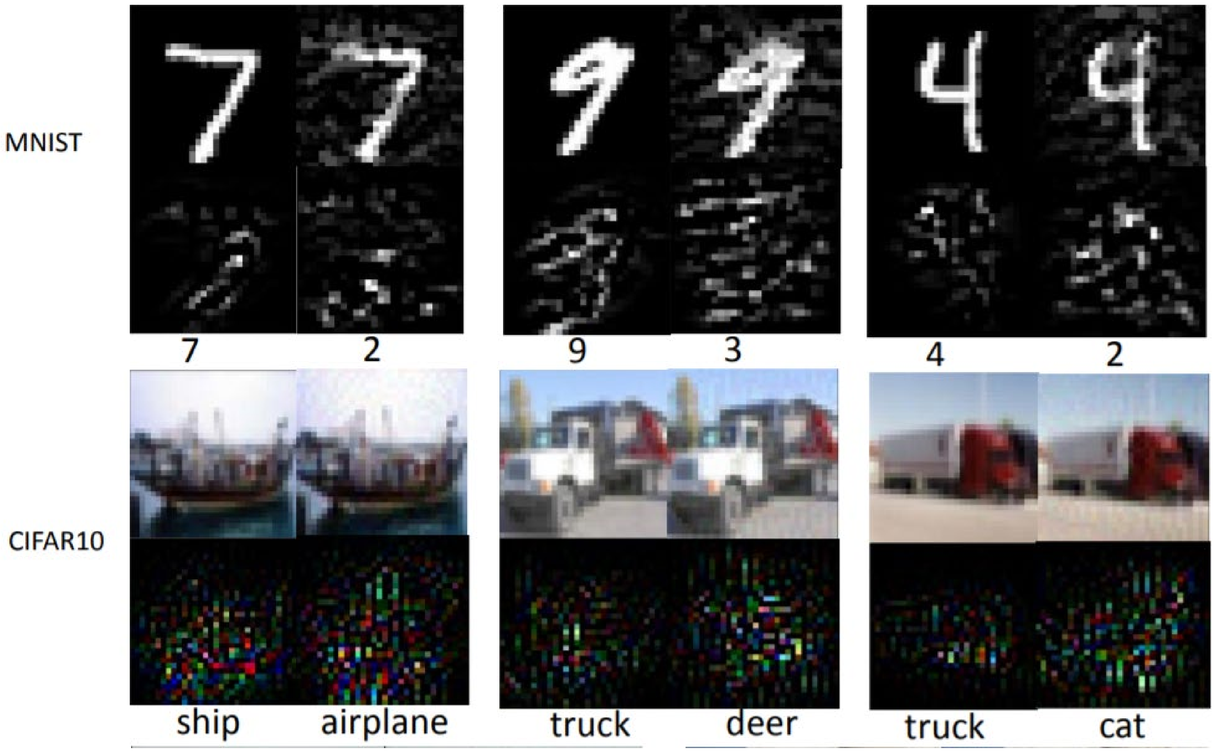
"Cat"

Other flavors

Method	What it does
GradCAM	Weight the 2D activations by the average gradient
GradCAM++	Like GradCAM but uses second order gradients
XGradCAM	Like GradCAM but scale the gradients by the normalized activations
AblationCAM	Zero out activations and measure how the output drops (this repository includes a fast batched implementation)
ScoreCAM	Perbutate the image by the scaled activations and measure how the output drops
EigenCAM	Takes the first principle component of the 2D Activations (no class discrimination, but seems to give great results)
EigenGradCAM	Like EigenCAM but with class discrimination: First principle component of Activations*Grad. Looks like GradCAM, but cleaner

https://github.com/jacobgil/pytorch-grad-cam

Saliency map for adversarial examples



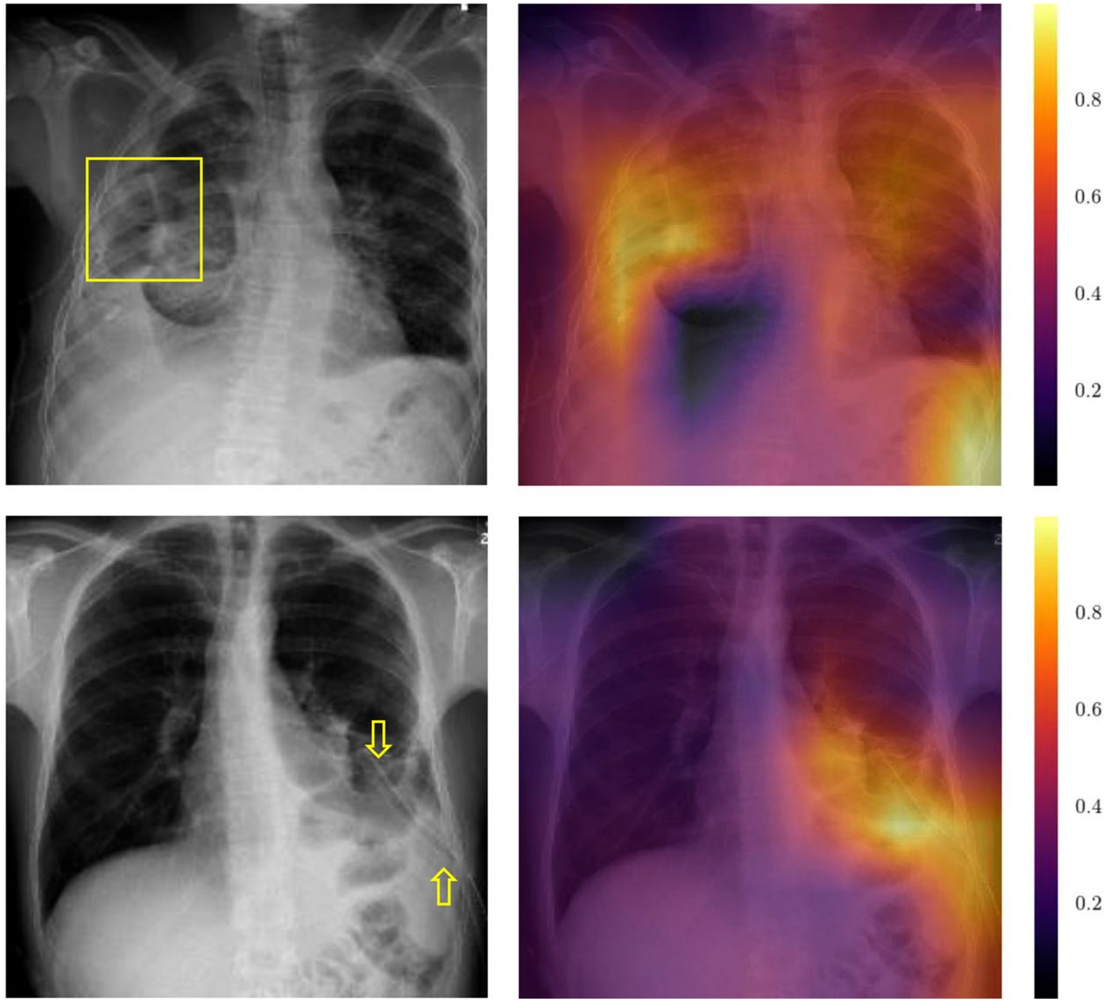
Train a binary classifier to take both image and saliency map as inputs

100% accuracy on detecting adversarial perturbations on MNIST dataset and show above 90% accuracy on CIFAR10.

- Saliency map is derived from gradient

Detecting Adversarial Perturbations with Saliency. https://arxiv.org/pdf/1803.08773.pdf

Saliency map can be useful

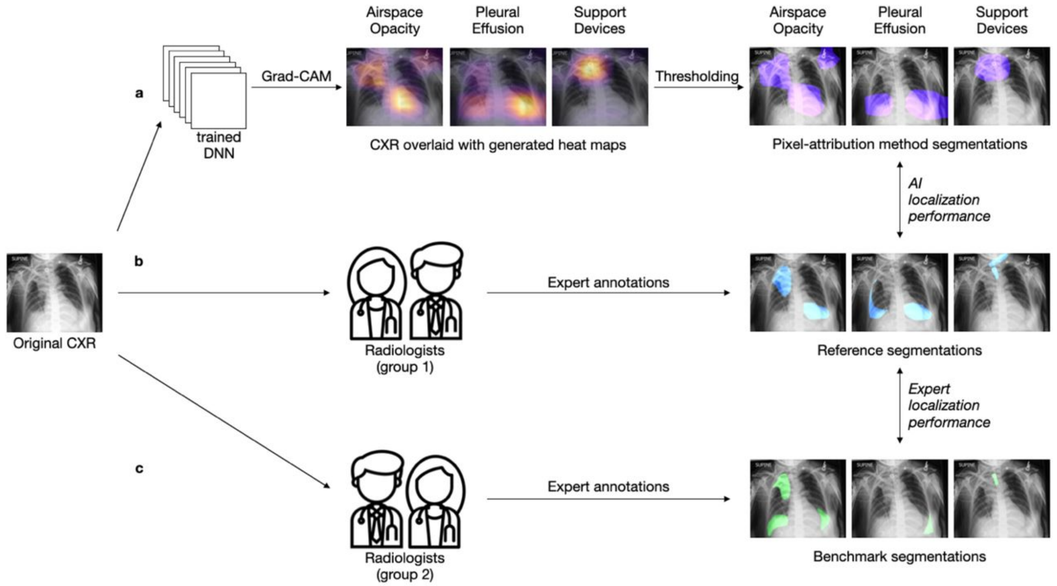


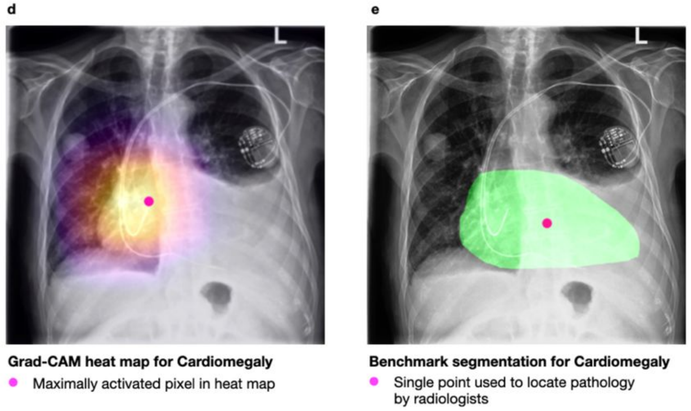
Positive pneumonia

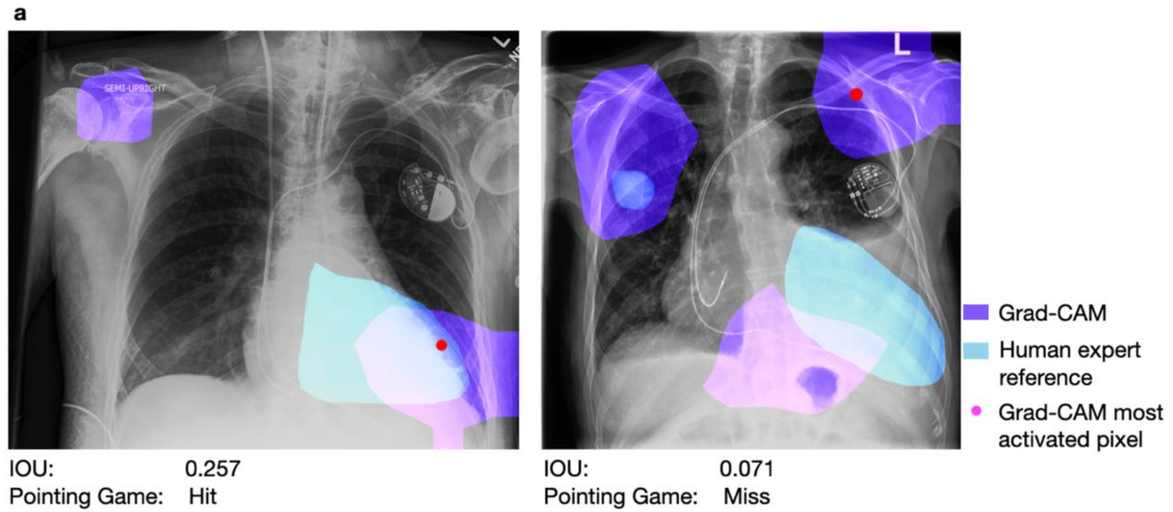
False positive, recognize drainpipe on patient chest x-ray as hotspot

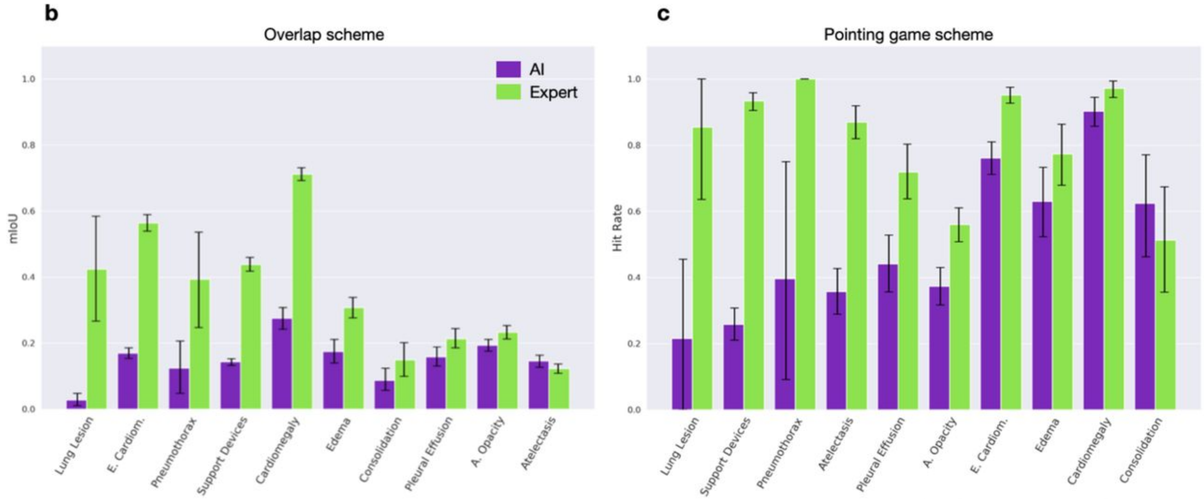
Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification. Scientific Reports volume 9, Article number: 6381 (2019)

But use it with care





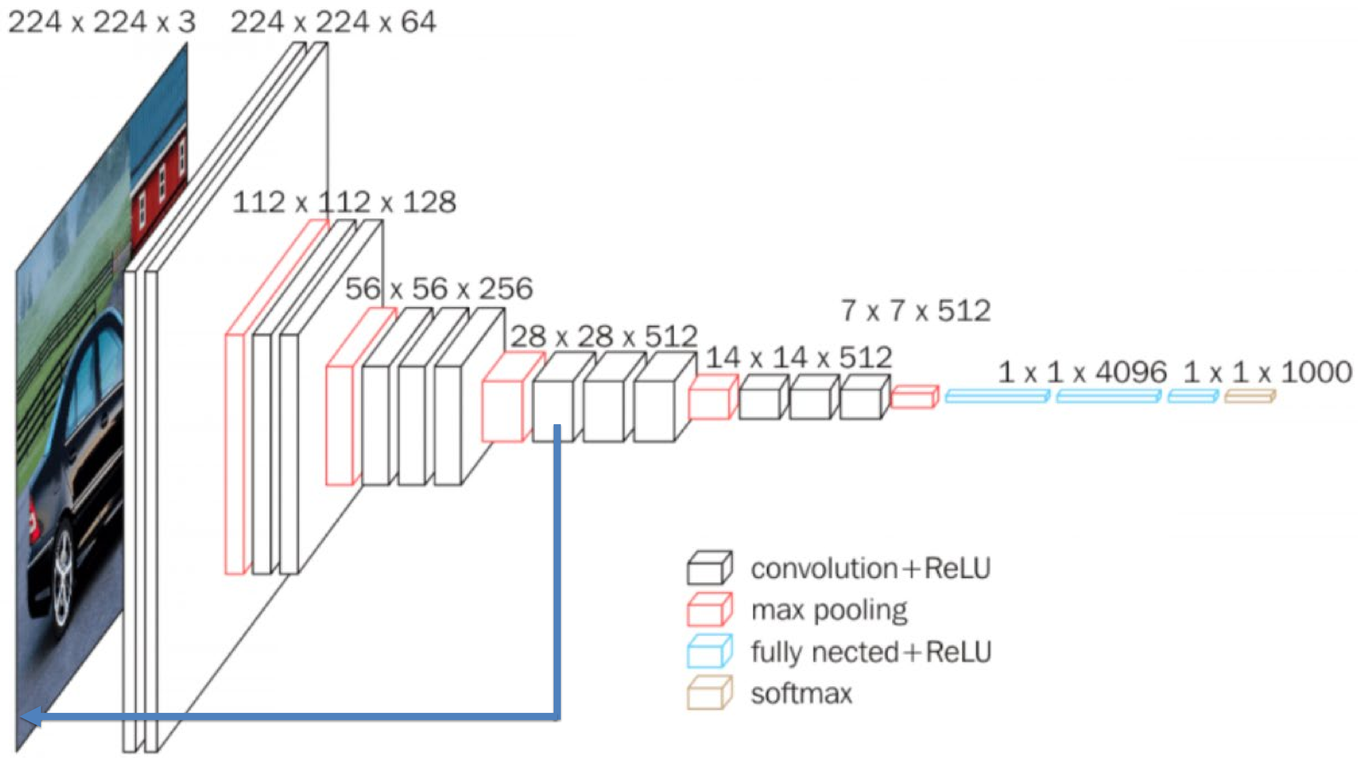




Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. https://www.medrxiv.org/content/10.1101/2021.02.28.21252634v1.full. March. 2021

National Heart, Lung, and Blood Institute

Guided backprop



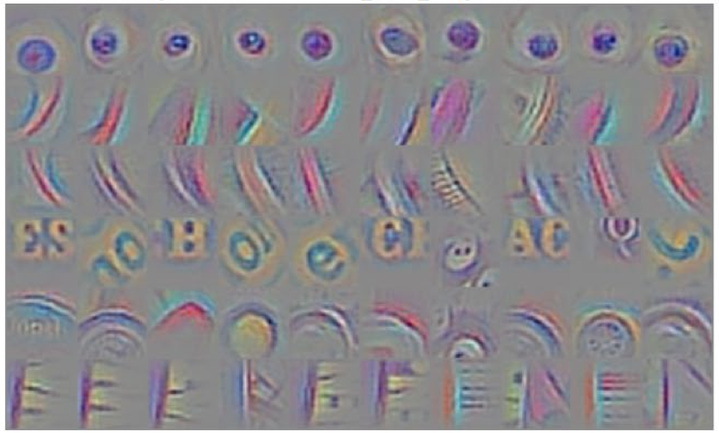
- Take one activation from a CONV layer, e.g. one value in 28x28x512 activation array
- Compute its gradient to input image x
- Gradient magnitude shows the part of image content activating this neuron

Very Deep Convolutional Networks for Large-Scale Image Recognition. VGG 16

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034v2

Guided backprop

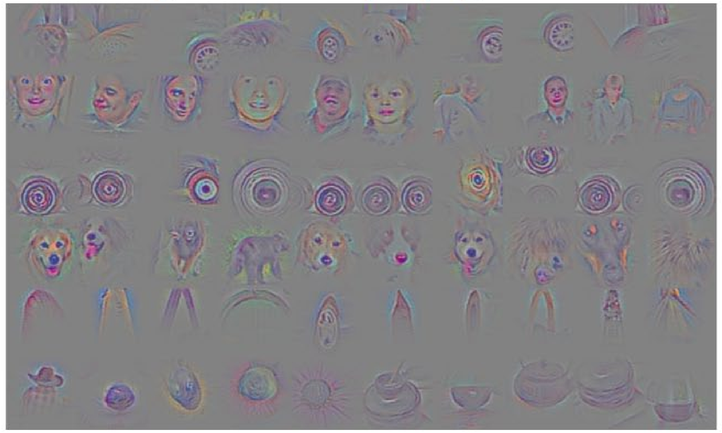
guided backpropagation



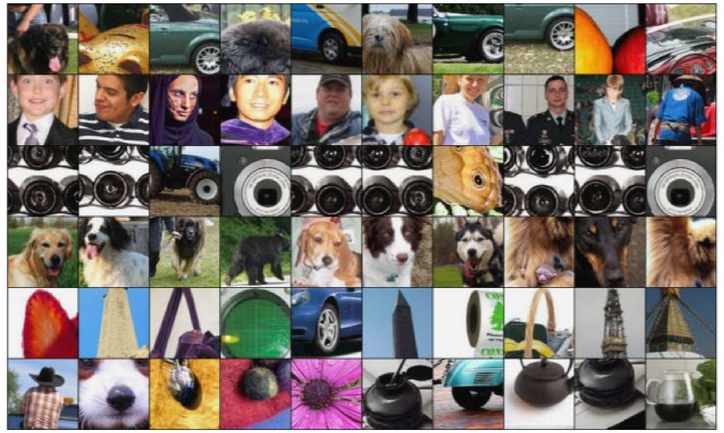
corresponding image crops

conv6

guided backpropagation



corresponding image crops



conv9

top 10 image patches activating this filter

Striving for Simplicity: The All Convolutional Net. https://arxiv.org/abs/1412.6806

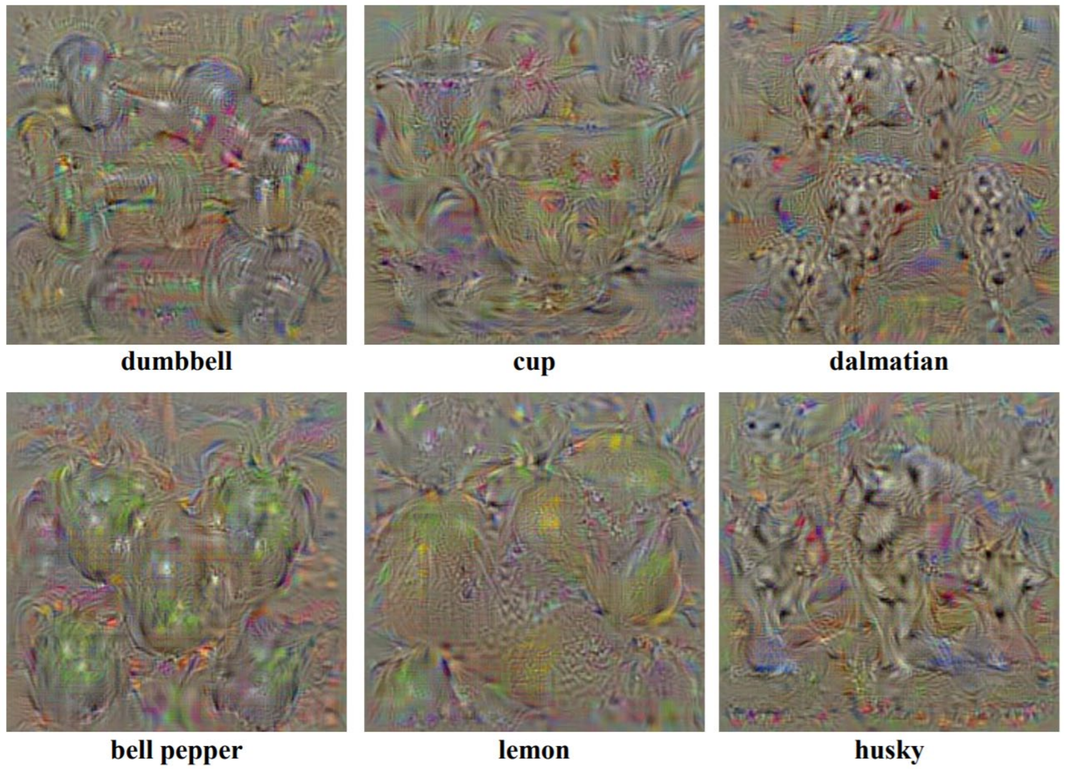
Gradient Ascending

- Global explanation of the trained model
- Generate an image to maximize a specific class core or loss

$$\arg\max_I S_c(I) - \lambda \|I\|_2^2$$

Class score

Regularization

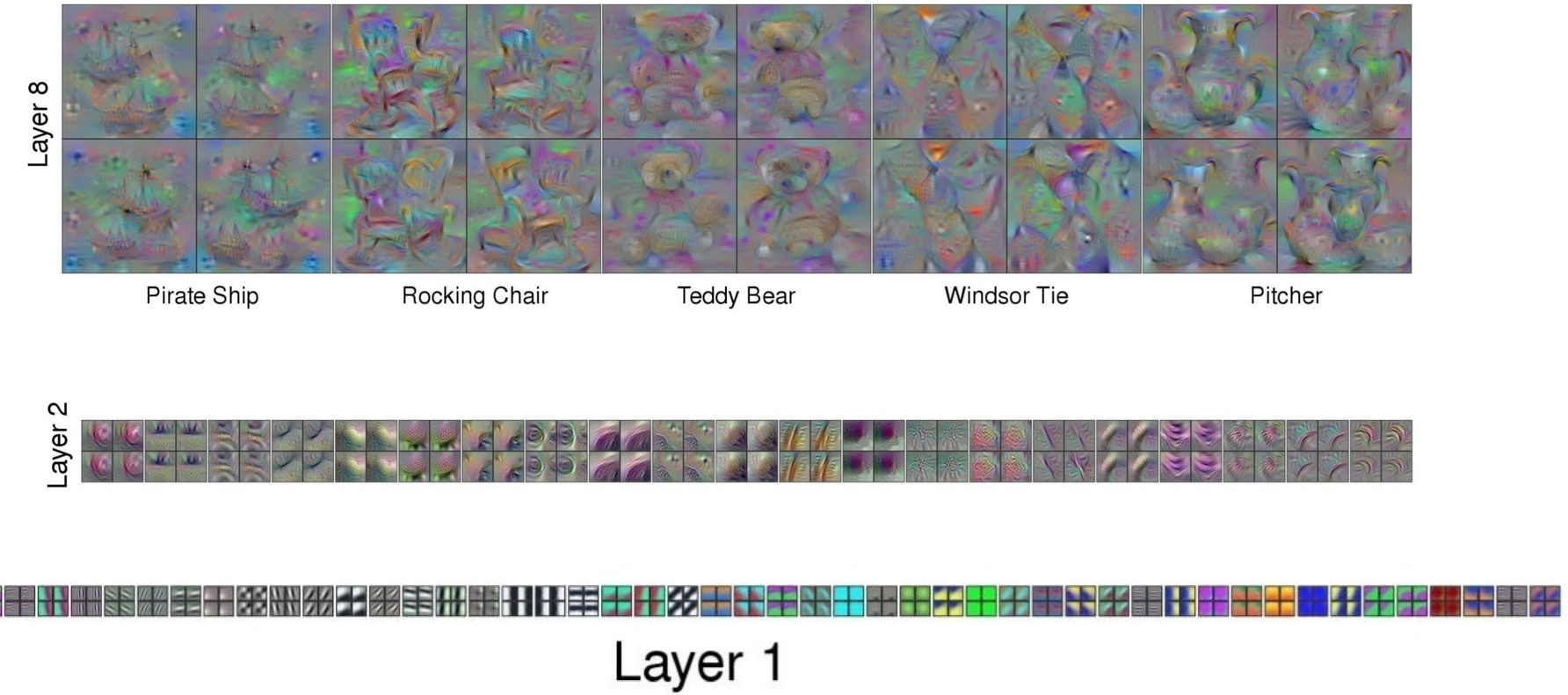


- Start from x=0
- Backprop to x
- Update x with gradient

Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. https://arxiv.org/abs/1312.6034

Gradient Ascending

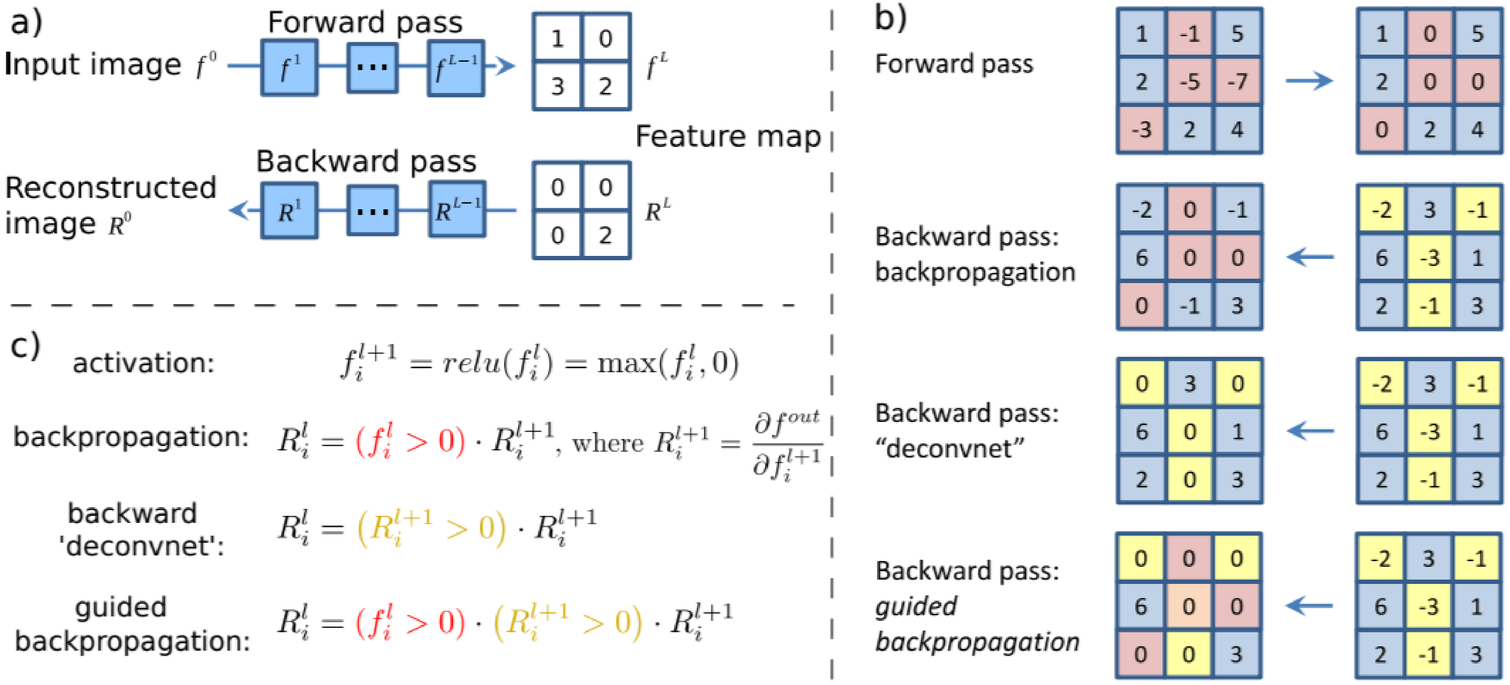
- Same scheme can be used to compute images to maximize an intermediate neuron



Understanding Neural Networks Through Deep Visualization. https://arxiv.org/abs/1506.06579

National Heart, Lung,
and Blood Institute

Guided backprop



- Improve the localization of class object
- Change is on ReLU layer
- Only backprop through ReLU layer if input value is positive and gradient is positive

Striving for Simplicity: The All Convolutional Net. https://arxiv.org/abs/1412.6806