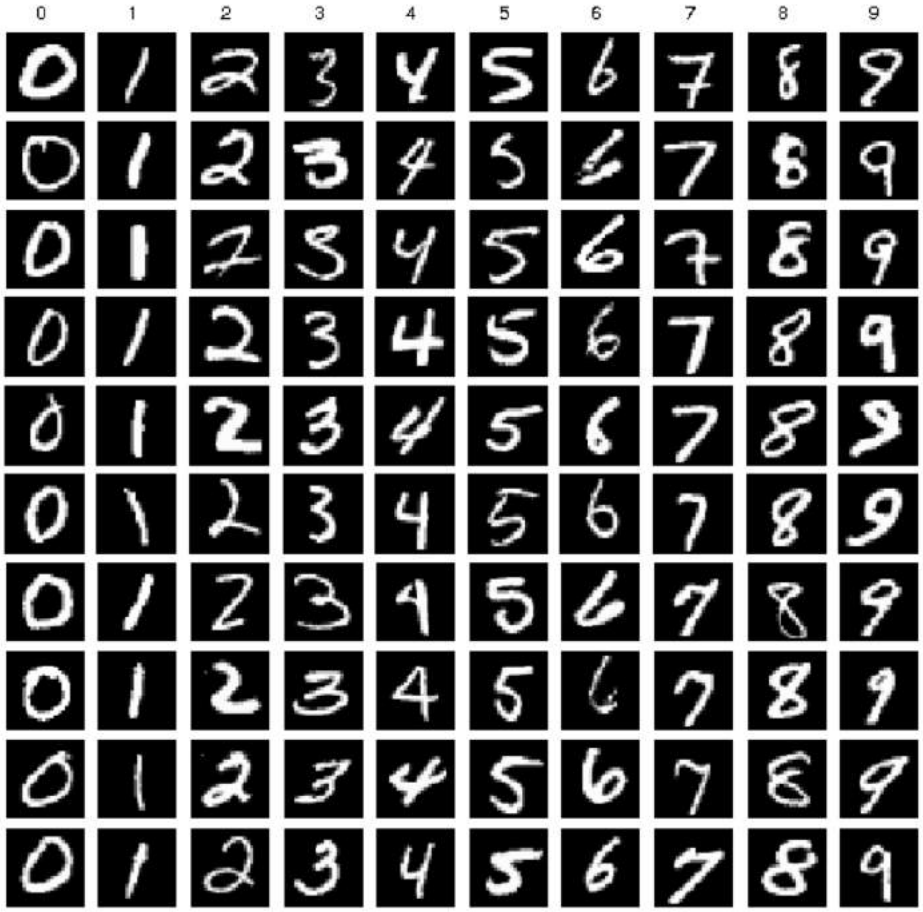
Image Classification

Image Classification

Which digit is presented in the image?

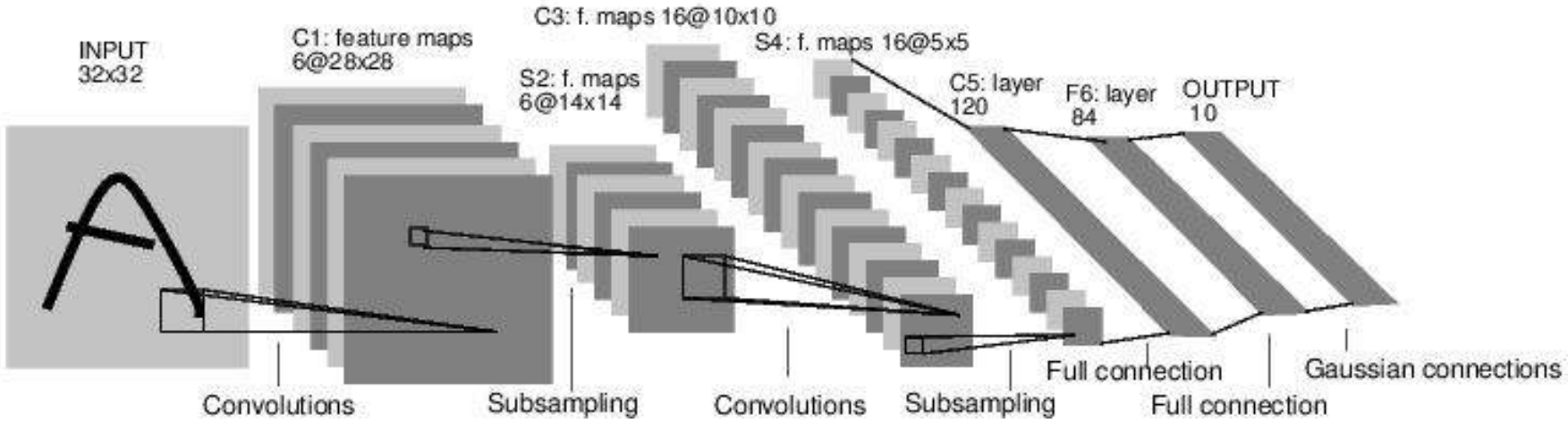


[LeCun, Cortes, Burges]

ConvNets for Image Classification

Case Study: LeNet-5

[LeCun et al., 1998]



Conv filters were 5x5, applied at stride 1
Subsampling (Pooling) layers were 2x2 applied at stride 2
i.e. architecture is [CONV-POOL-CONV-POOL-CONV-FC]

Image Classification

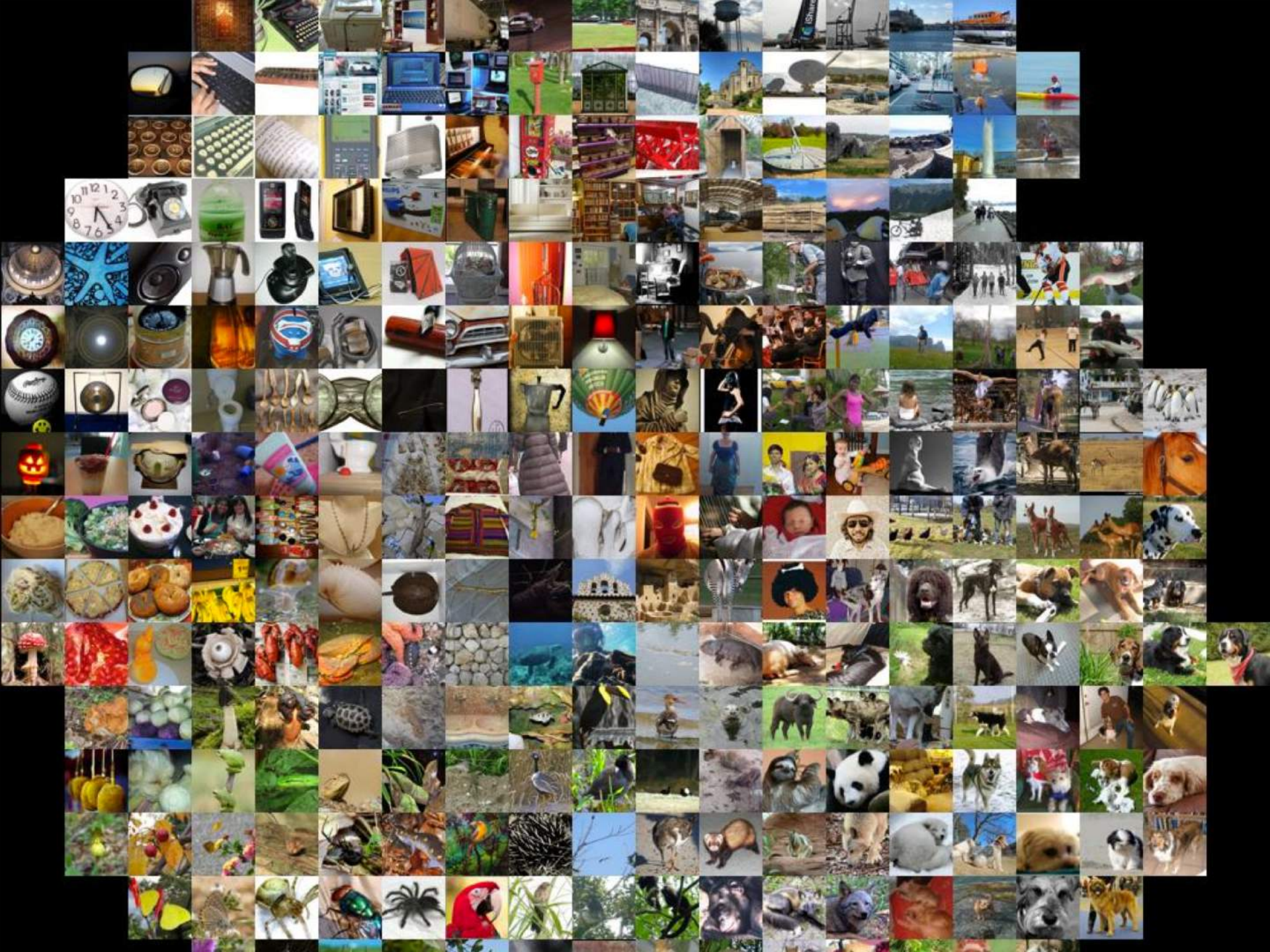
What object is presented in the image?

ImageNet: A Large-Scale Hierarchical Image Database,
Deng, Dong, Socher, Li, Li and Fei-Fei, *CVPR,* 2009

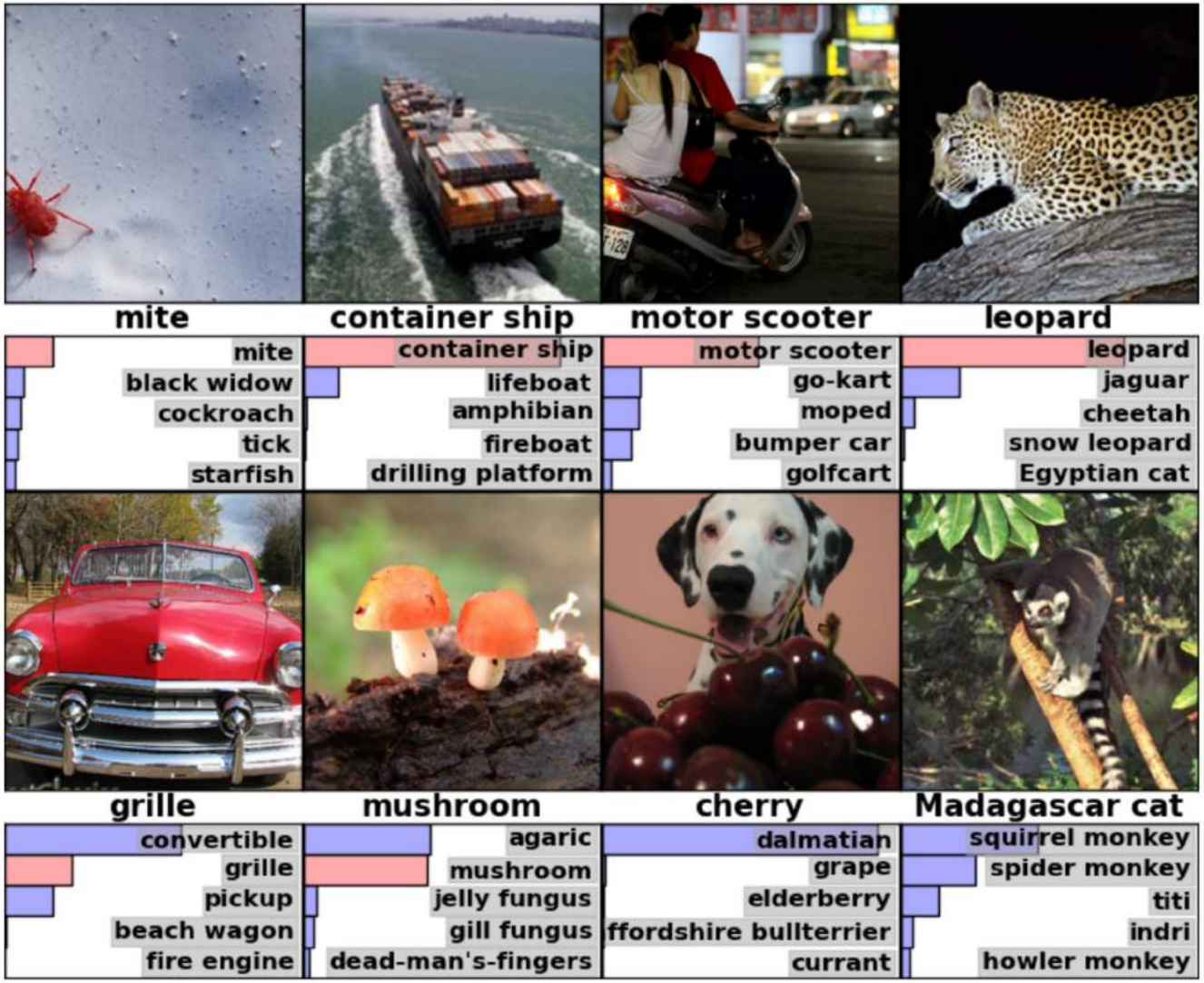
ImageNet Classification

What object is presented in the image?

- Assume a single object of interest
- Target object usually lies around the center
- Somewhat random set of objects (e.g., dogs)
- Still interesting and very challenging!

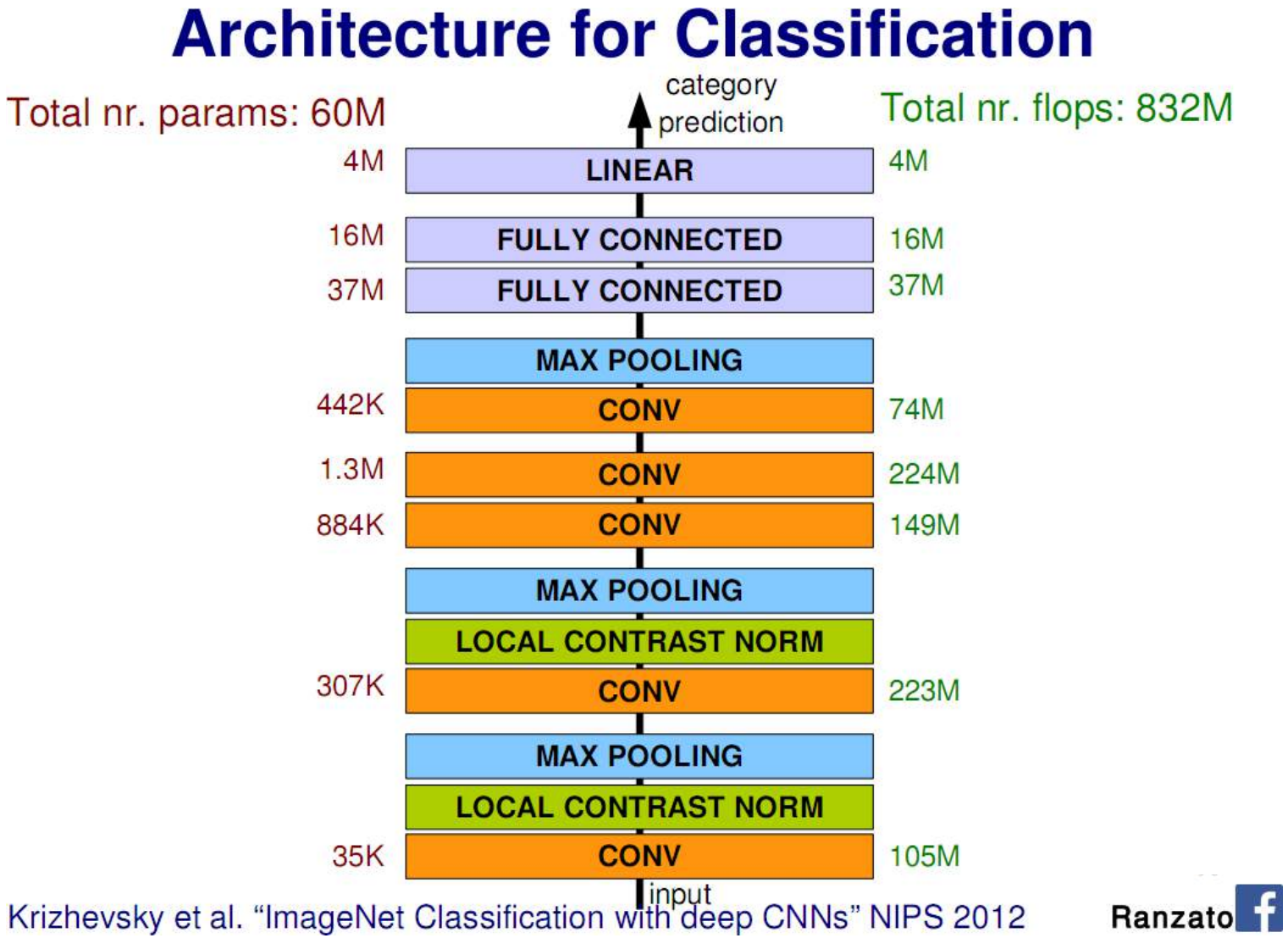


ConvNets for Image Classification



"ImageNet Classification with Deep Convolutional Neural Networks",
Krizhevsky, Sutskever, Hinton, *NIPS,* 2012

ConvNets for Image Classification



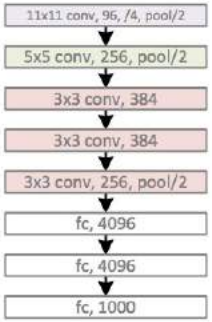
Krizhevsky et al. "ImageNet Classification with deep CNNs" NIPS 2012

Ranzato

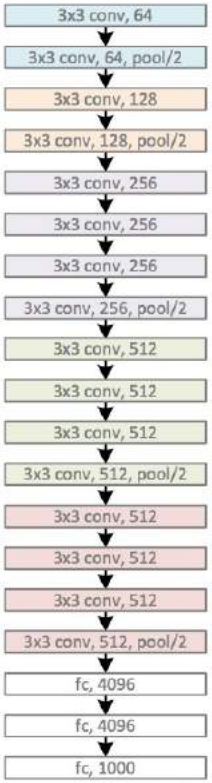
ConvNets for Image Classification

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)



VGG, 19 layers
(ILSVRC 2014)



GoogleNet, 22 layers
(ILSVRC 2014)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Slide Credit: Kaiming He

ConvNets for Image Classification

Revolution of Depth

AlexNet, 8 layers
(ILSVRC 2012)

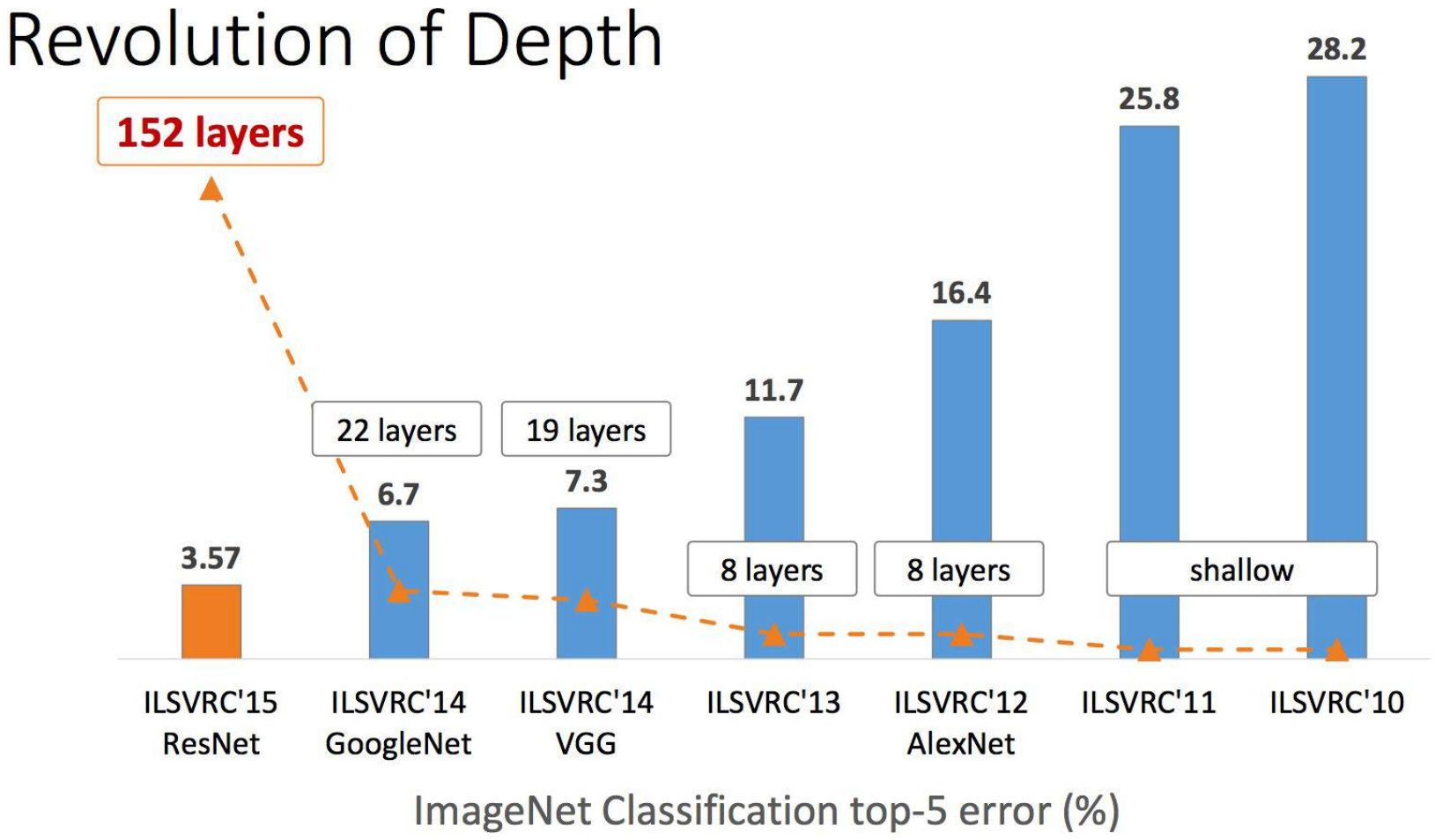
VGG, 19 layers
(ILSVRC 2014)

ResNet, 152 layers
(ILSVRC 2015)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". arXiv 2015.

Slide Credit: Kaiming He

ConvNets for Image Classification

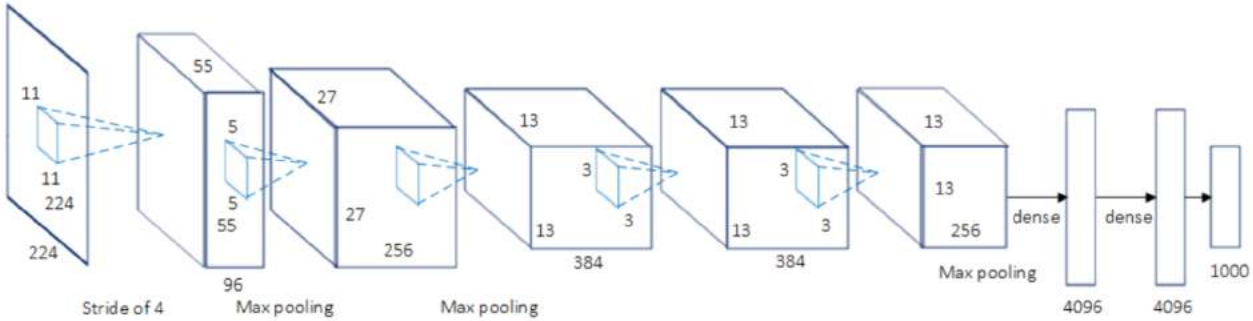


Slide Credit: Kaiming He

What do we know about these networks?

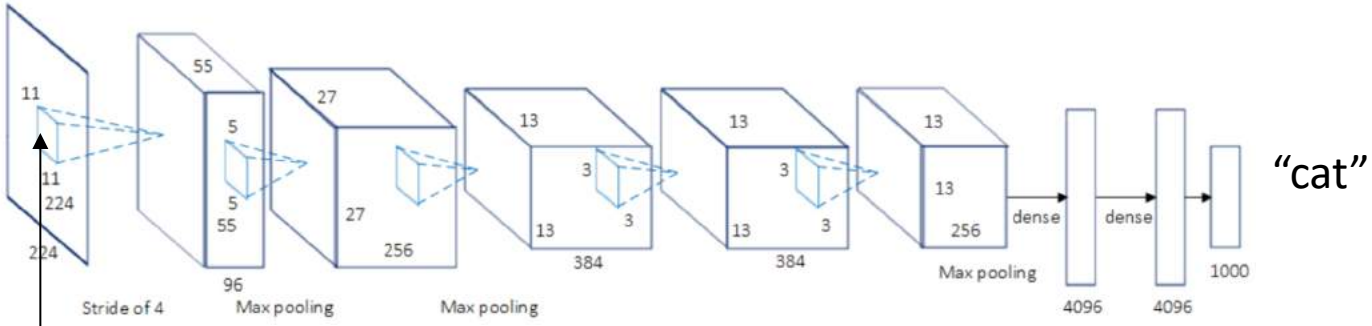
Many slides from Lana Lazebnik

What has been learned by ConvNets?

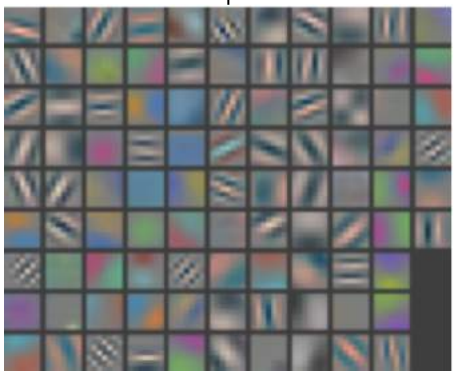


"cat"

What has been learned by ConvNets?



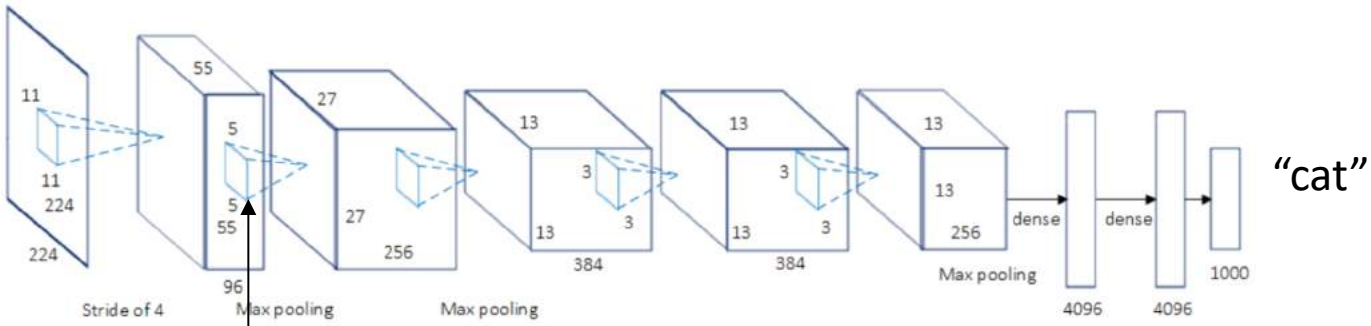
Visualize first-layer weights directly



Visualizing ConvNets

- “It’s a black box!”
- “Nobody understands what’s going on!”
- “Conv1 is gabor filters, but what’s actually going on?!”
- “Sure, LeCun and Hinton know how to make them work, but it’s magic.”

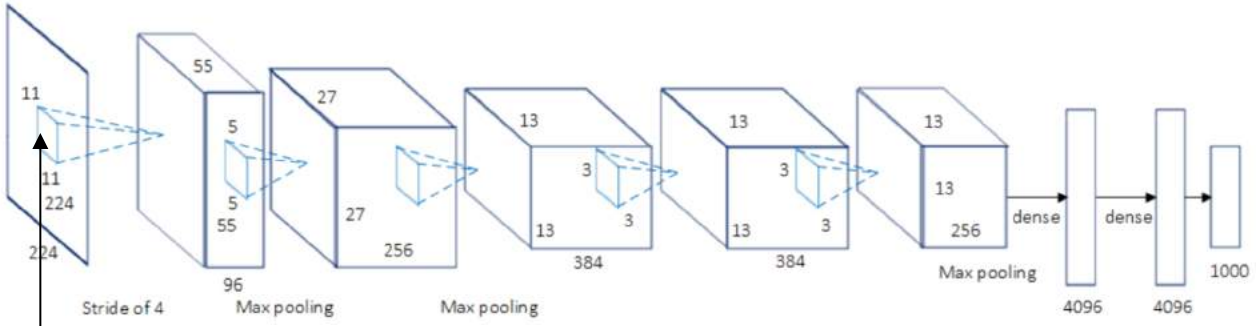
What has been learned by ConvNets?



Not too helpful for subsequent layers

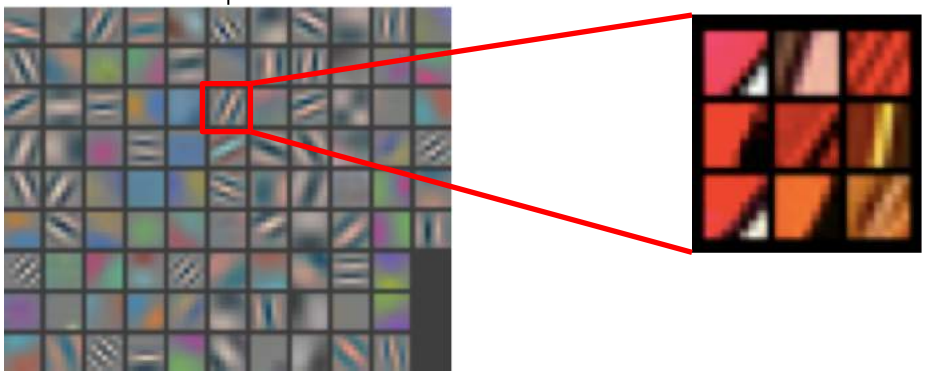
Features from a CIFAR10 network, via Stanford CS231n

What has been learned by ConvNets?

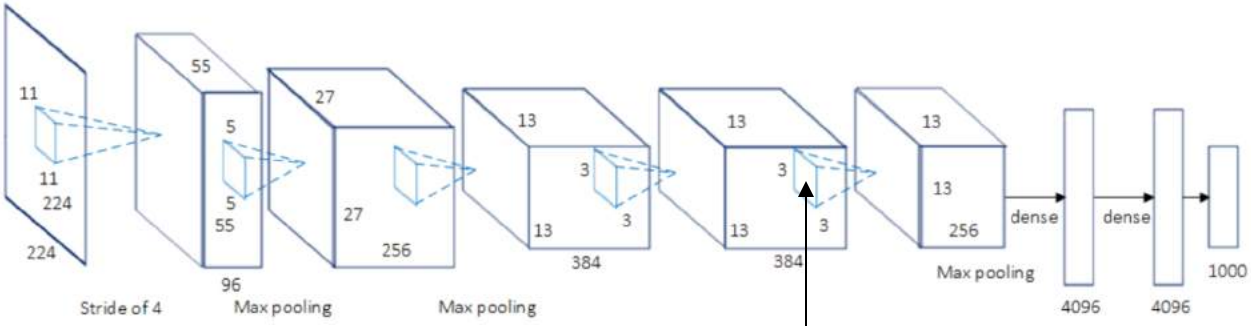


Visualize maximally activating patches:

pick a unit; run many images through the network;
visualize patches that produce the highest output values

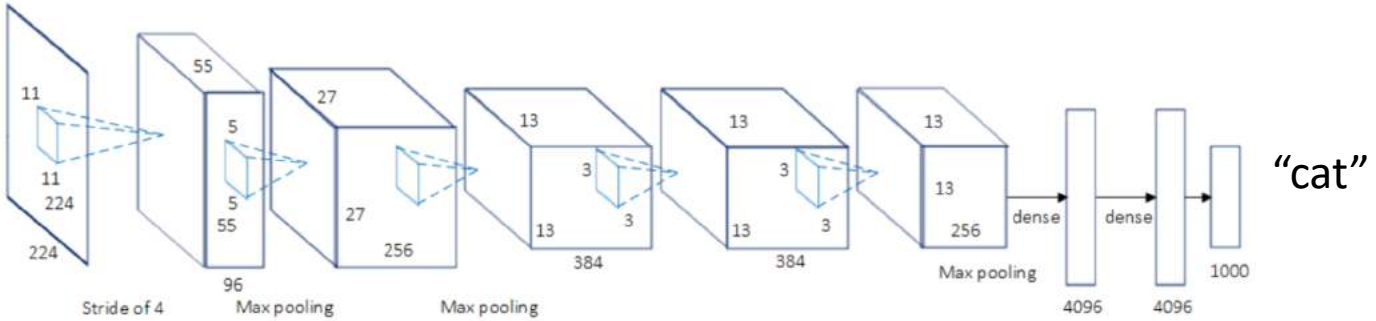


What has been learned by ConvNets?



Visualize maximally activating patches

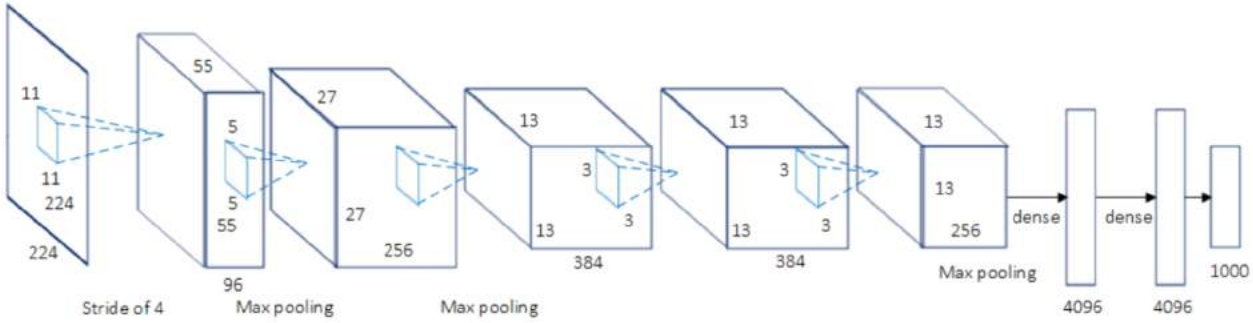
What has been learned by ConvNets?

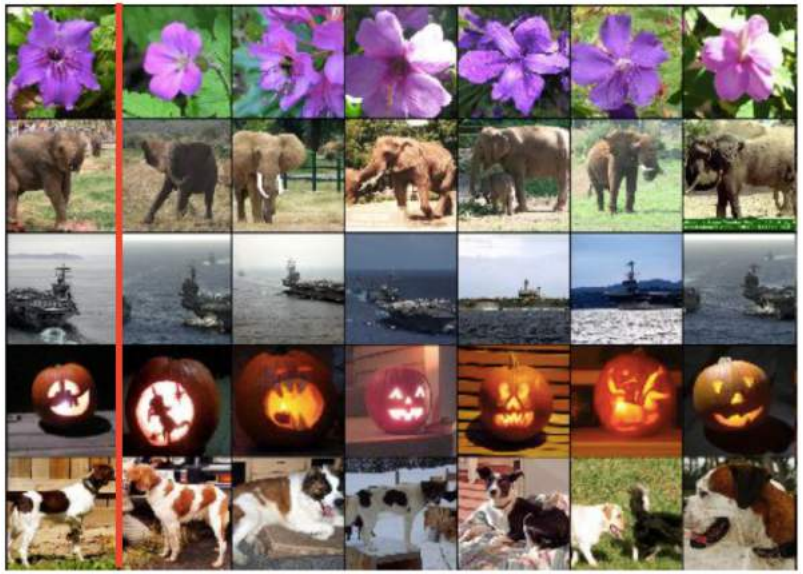


A tour through the network!

Visualization: Thanks to David Fouhey

What has been learned by ConvNets?



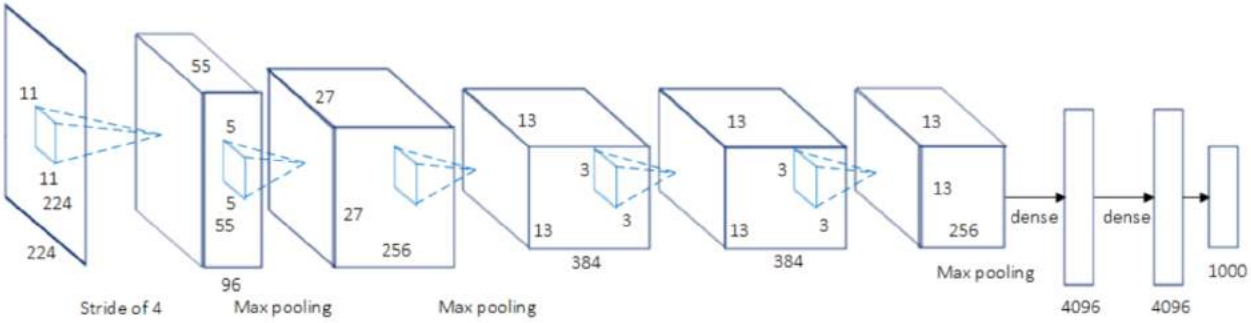


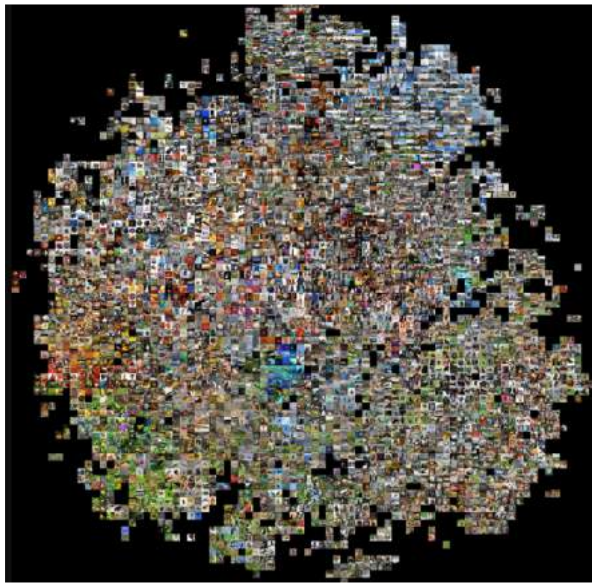
FC layers?

Visualize nearest neighbors according to activation vectors

Source: Stanford CS231n

What has been learned by ConvNets?



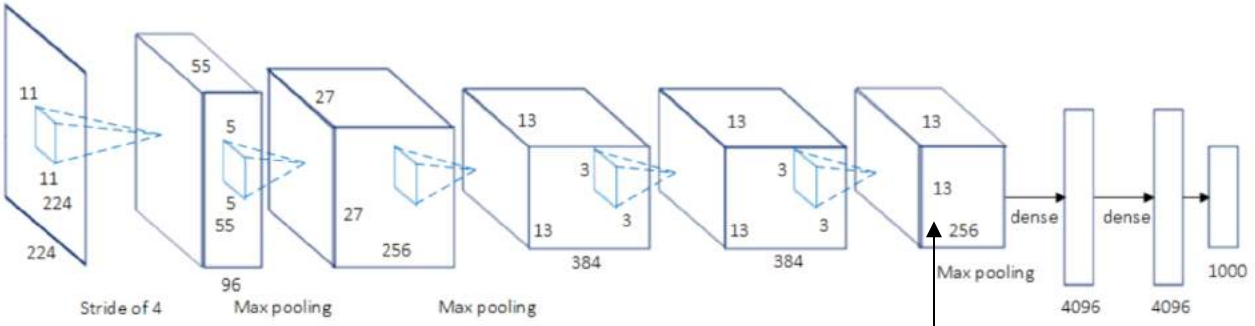


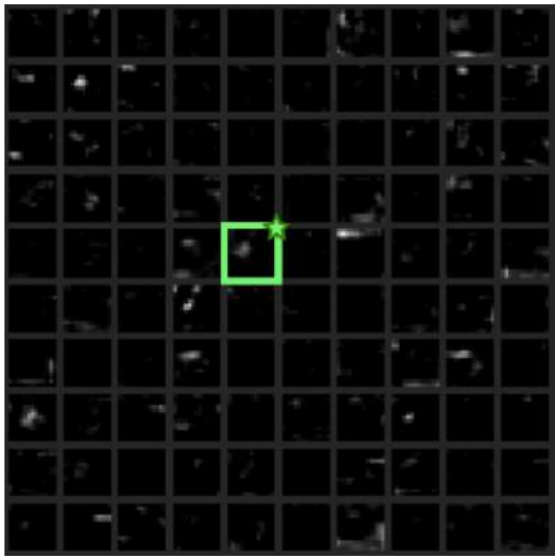
FC layers?

Fancy dimensionality reduction, e.g., t-SNE

Source: Andrej Karpathy

What has been learned by ConvNets?

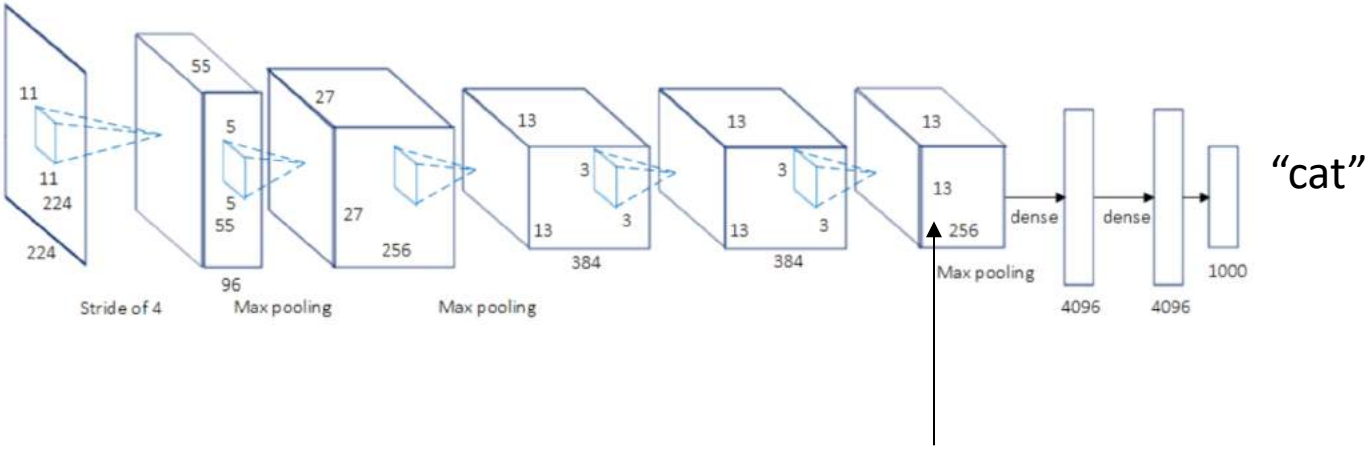




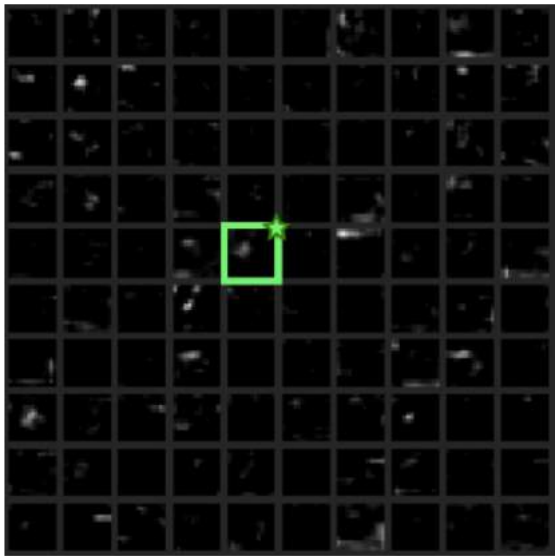
Visualize activations for an image

Source

What has been learned by ConvNets?



Going back to image?

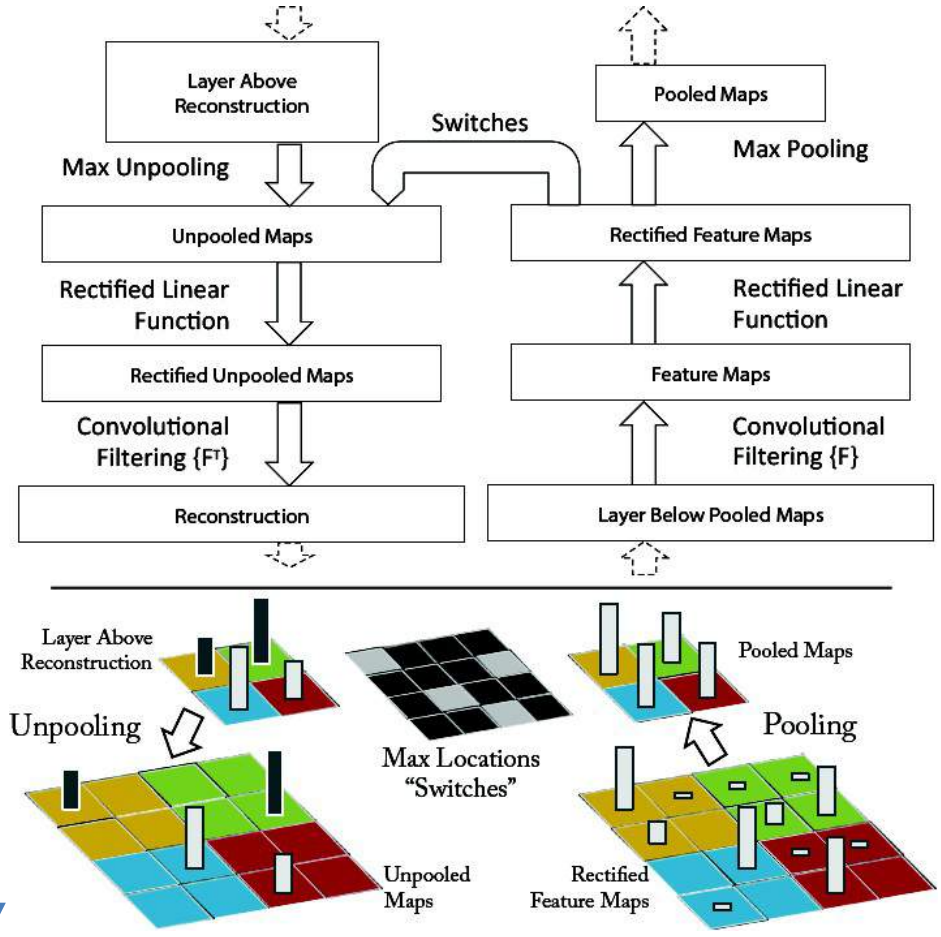


Visualize activations for an image

Source

Visualizing ConvNets: Going back to images

Towards Image



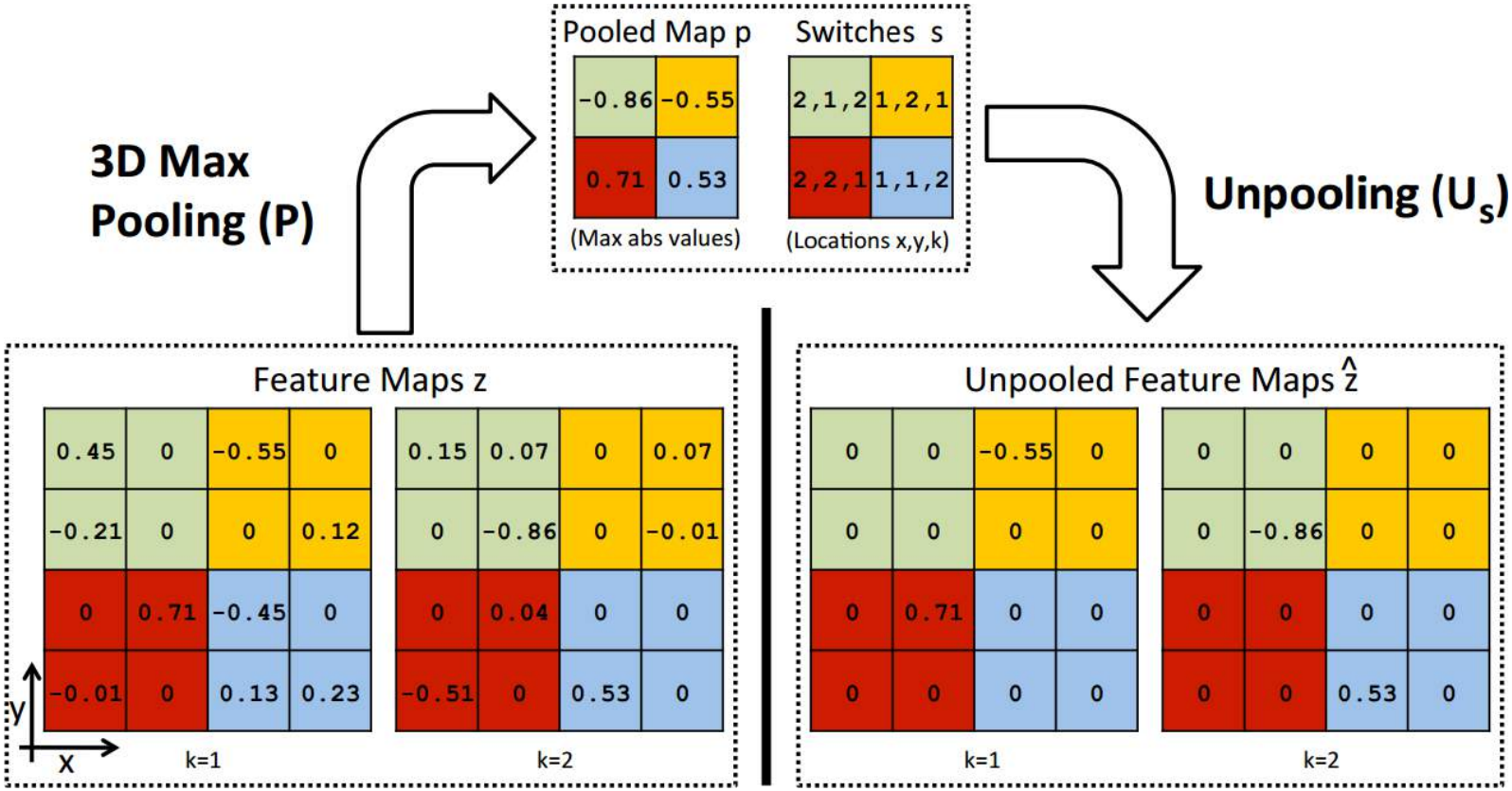
Towards Predictions

[Zeiler and Fergus]

Visualizing ConvNets: Things to invert

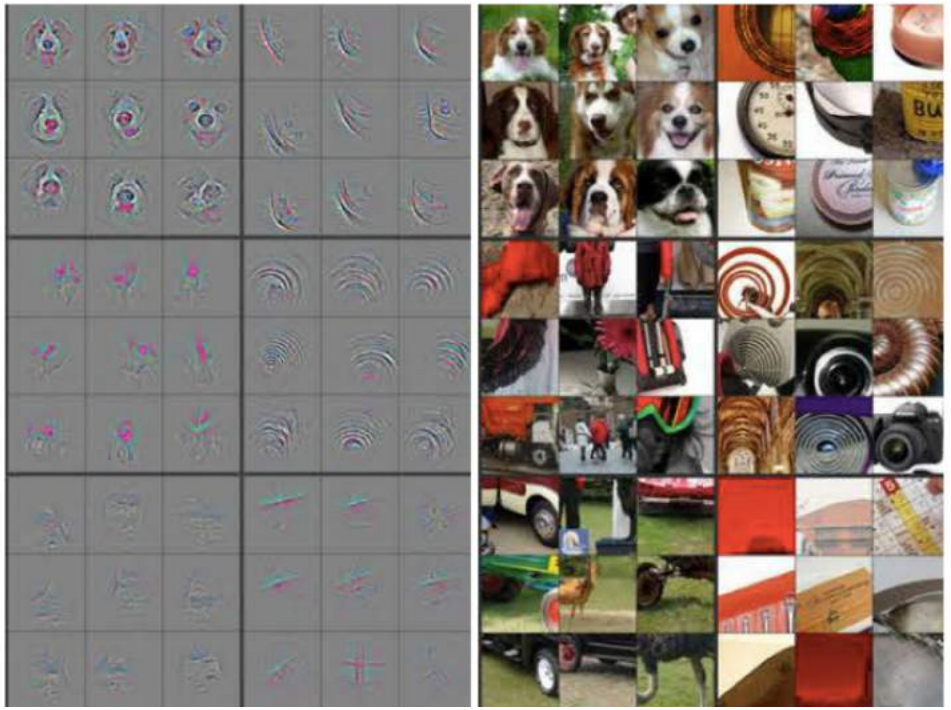
- Convolutions/Filtering
- Rectification/Non-linearity
- Pooling

Visualizing ConvNets: Things to invert

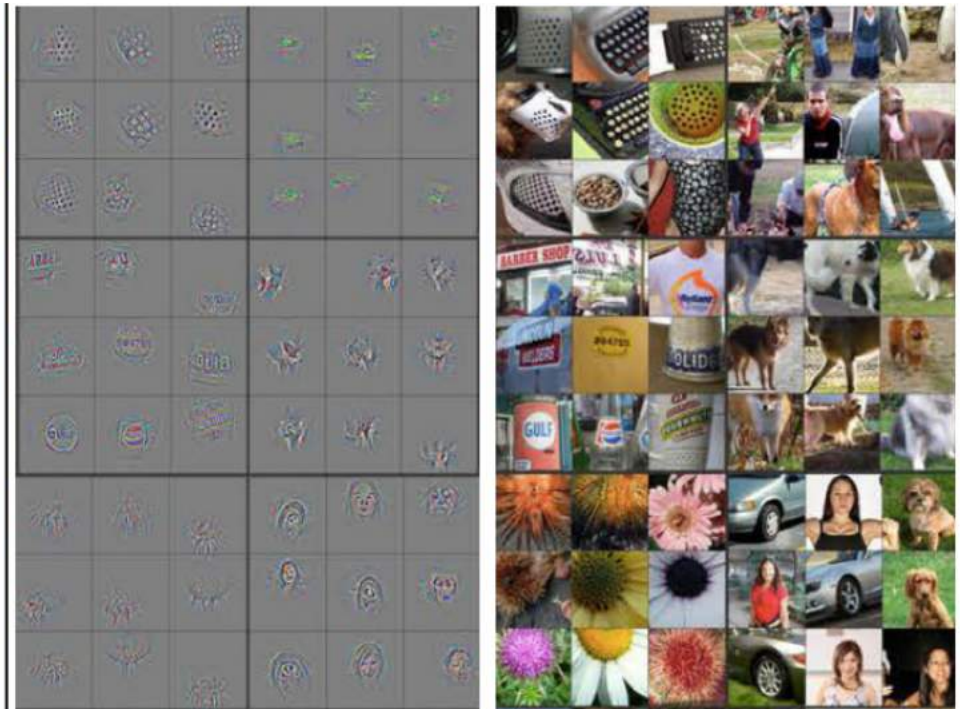


Visualizing ConvNets: Going back to images

AlexNet Layer 4



AlexNet Layer 5



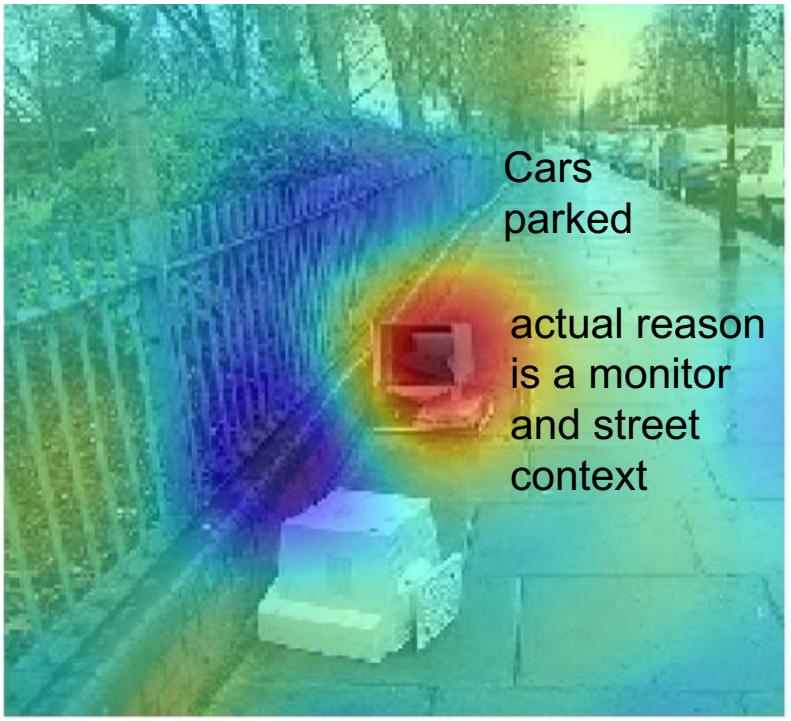
M. Zeiler and R. Fergus,
Visualizing and Understanding Convolutional Networks, ECCV 2014

Visualizing ConvNets: Saliency Maps

Which parts of the image played the most important role in the network's decision?

Prediction: "car" 64%

Explanation for "car"



Source: K. Saenko

“White Box” Saliency: Using Gradients

Compute $-\frac{\partial L(x;\theta)}{x}$ and display the max of absolution values across three channels

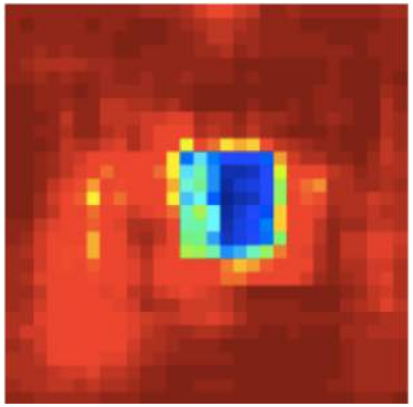
K. Simonyan, A. Vedaldi, and A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, ICLR 2014

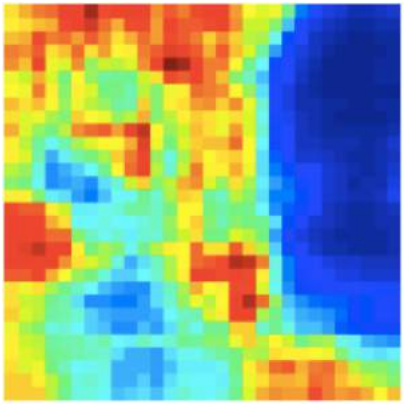
“Black Box” Saliency: Via Masking

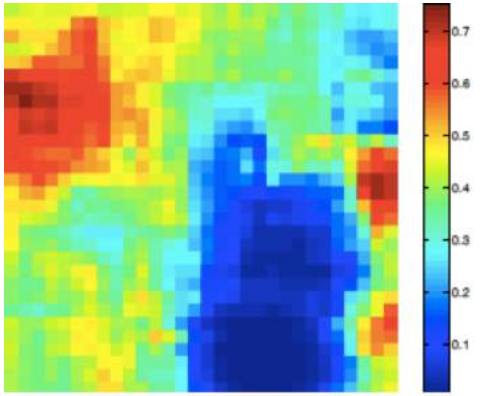
Slide square occluder across image, see how class score changes

Input image

Correct class probability as function of occluder position



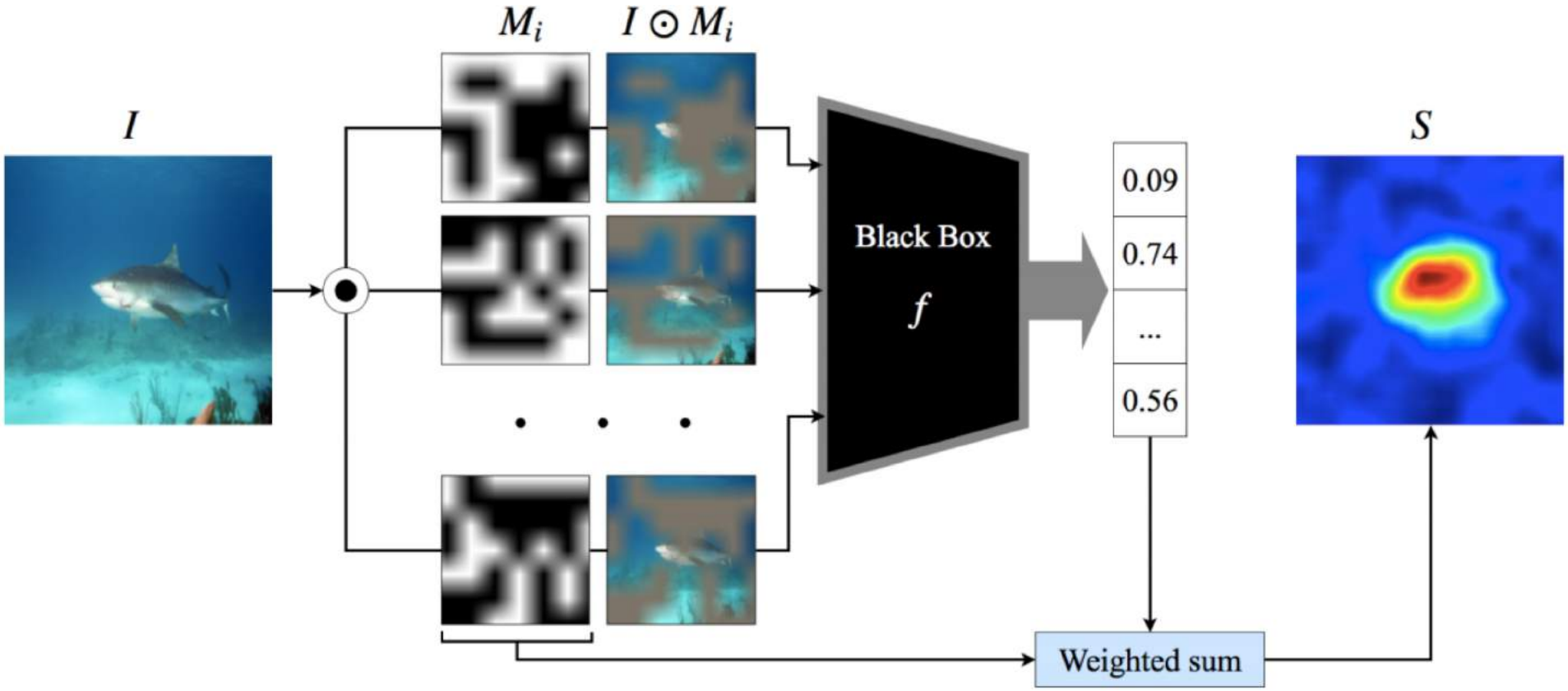




M. Zeiler and R. Fergus, Visualizing and Understanding Convolutional Networks, ECCV 2014

“Black Box” Saliency: Via Masking

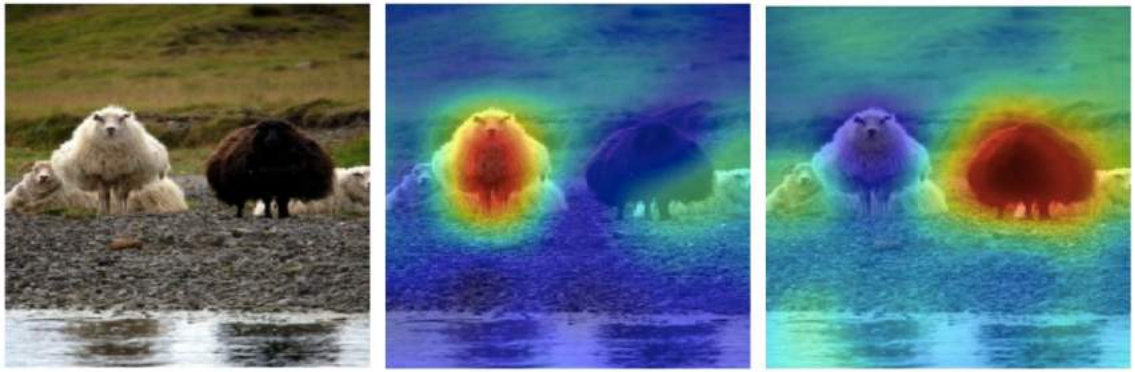
Pixel saliency is averaged across all masks



V. Petsiuk, A. Das, K. Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models, BMVC 2018

“Black Box” Saliency: Via Masking

Pixel saliency is class dependent



(a) Sheep - 26%, Cow - 17% (b) Importance map of ‘*sheep*’ (c) Importance map of ‘*cow*’

(d) Bird - 100%, Person - 39% (e) Importance map of ‘*bird*’ (f) Importance map of ‘*person*’

V. Petsiuk, A. Das, K. Saenko, RISE: Randomized Input Sampling for Explanation of Black-box Models, BMVC 2018

Quantify Interpretability of Units

- From the very beginning, people have observed that many units in higher layers seem to fire on semantically meaningful concepts
- But how can we quantify this?

AlexNet Layer 4

AlexNet Layer 5

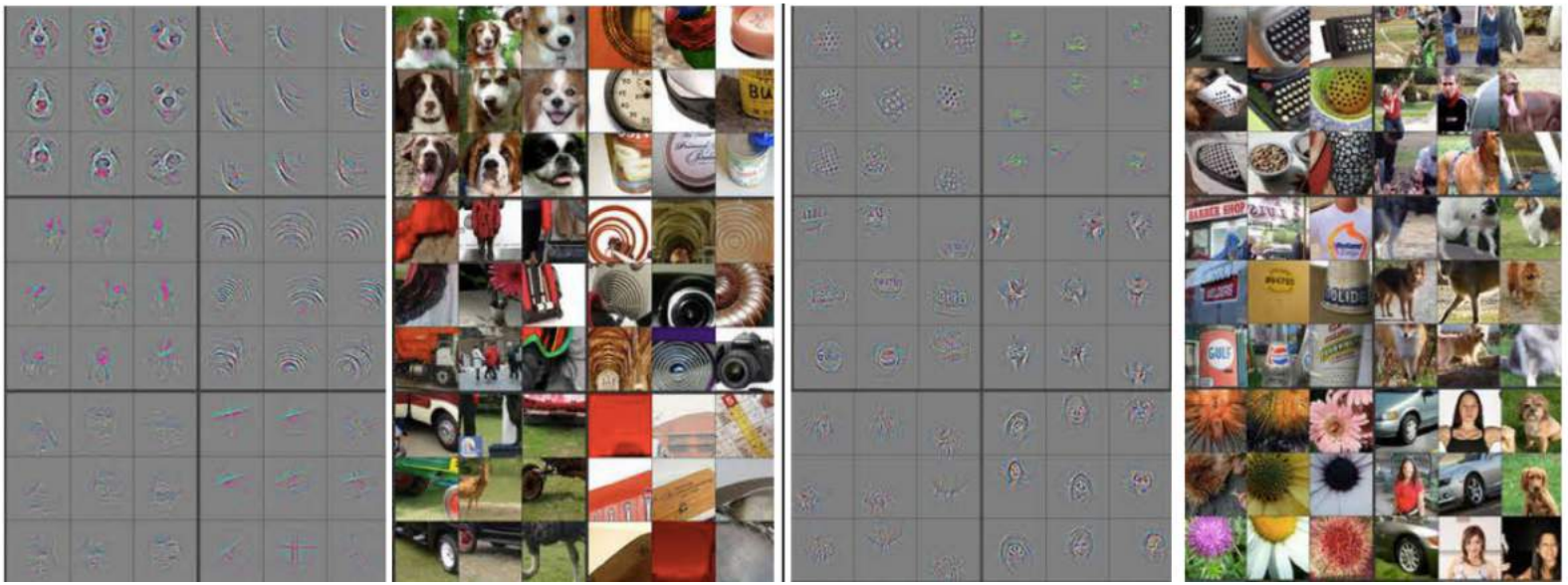
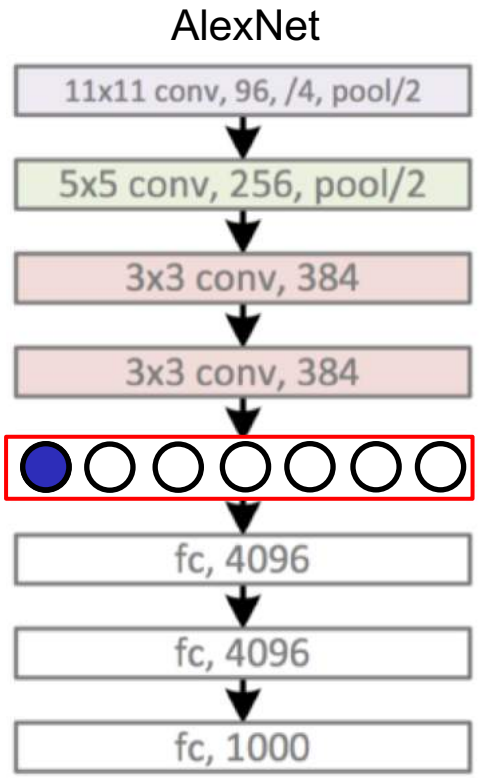
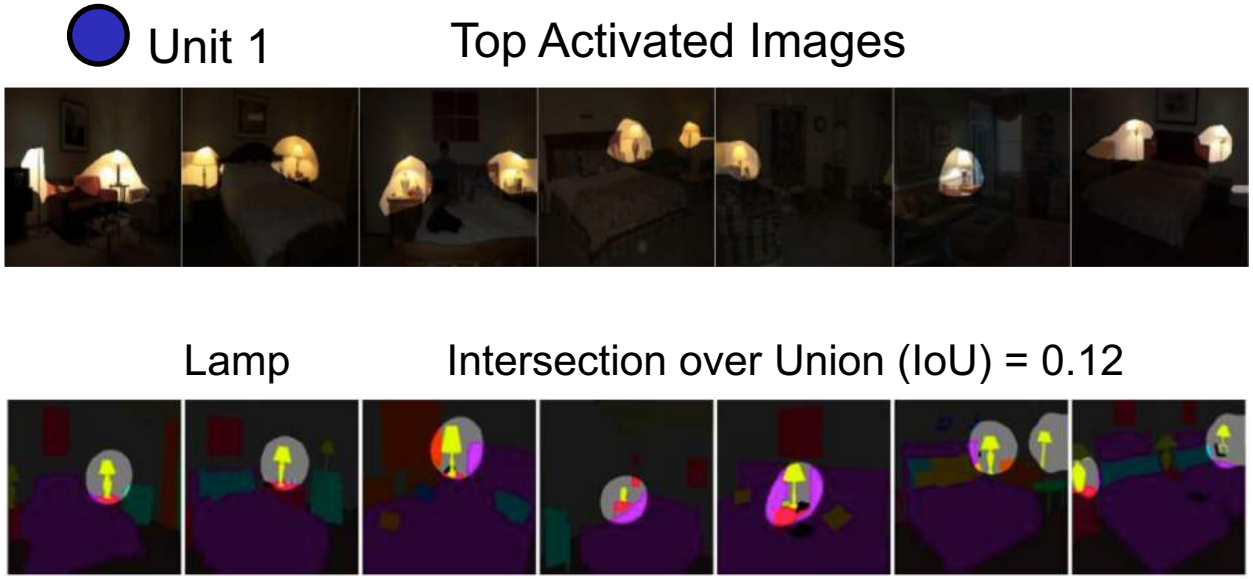


Figure: Zeiler & Fergus

Quantify Interpretability of Units

For a given unit, measure how much areas of high activation overlap semantic segmentations for a large set of visual concepts

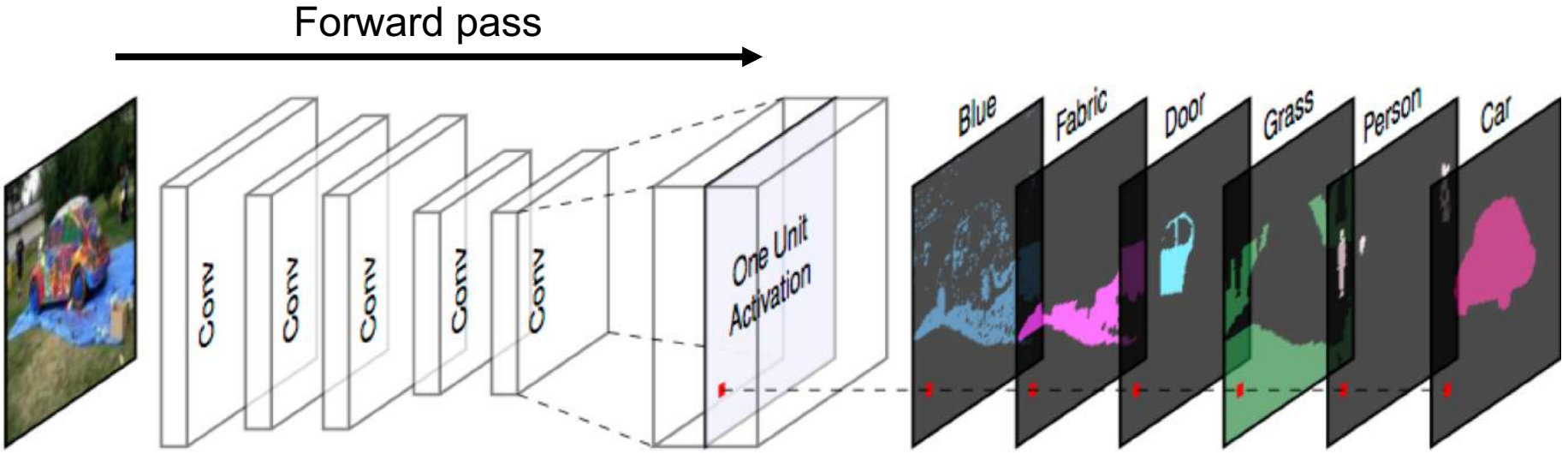




D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network Dissection: Quantifying Interpretability of Deep Visual Representations, CVPR 2017

Quantify Interpretability of Units

Assumption: grandmother cells instead of population coding in high parts of the network (semantic concept is represented by a single unit)



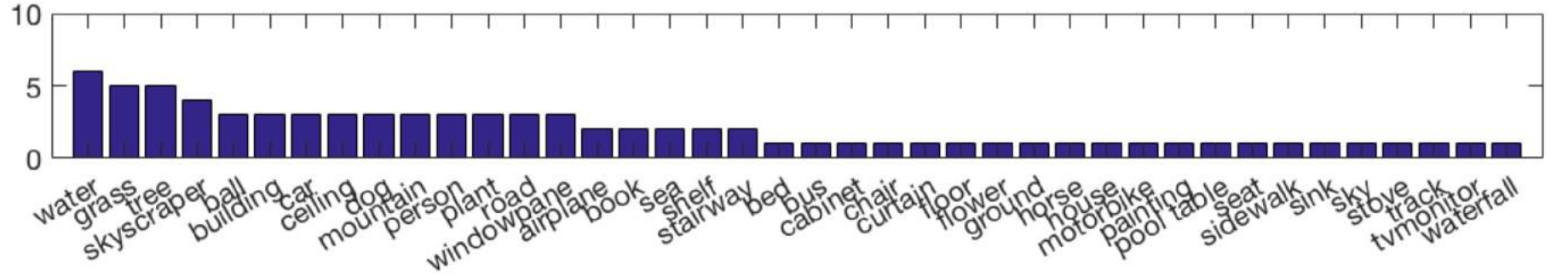
Network trained for image classification on ImageNet or Places

Dataset with semantic segmentations for ~1200 visual concepts

D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network Dissection: Quantifying Interpretability of Deep Visual Representations, CVPR 2017

Quantify Interpretability of Units

Histogram of object detectors for Places AlexNet conv5 units
81/256 units with IoU > 0.04



conv5 unit 79 car (object) IoU=0.13

conv5 unit 107 road (object) IoU=0.15

D. Bau, B. Zhou, A. Khosla, A. Oliva, A. Torralba, Network Dissection: Quantifying Interpretability of Deep Visual Representations, CVPR 2017

ConvNet Visualizations: Recap

- Basic visualization techniques
  - showing weights
  - top activated patches
  - nearest neighbors
- Mapping activations back to the image
  - Deconvolution
- Saliency maps
  - “White box” vs. “black box”
- Explainability/interpretability
  - Explaining network decisions
  - Quantifying interpretability of intermediate units

Are these networks robust?

Many slides from Lana Lazebnik

Fooling ConvNets

$+ .007 \times$

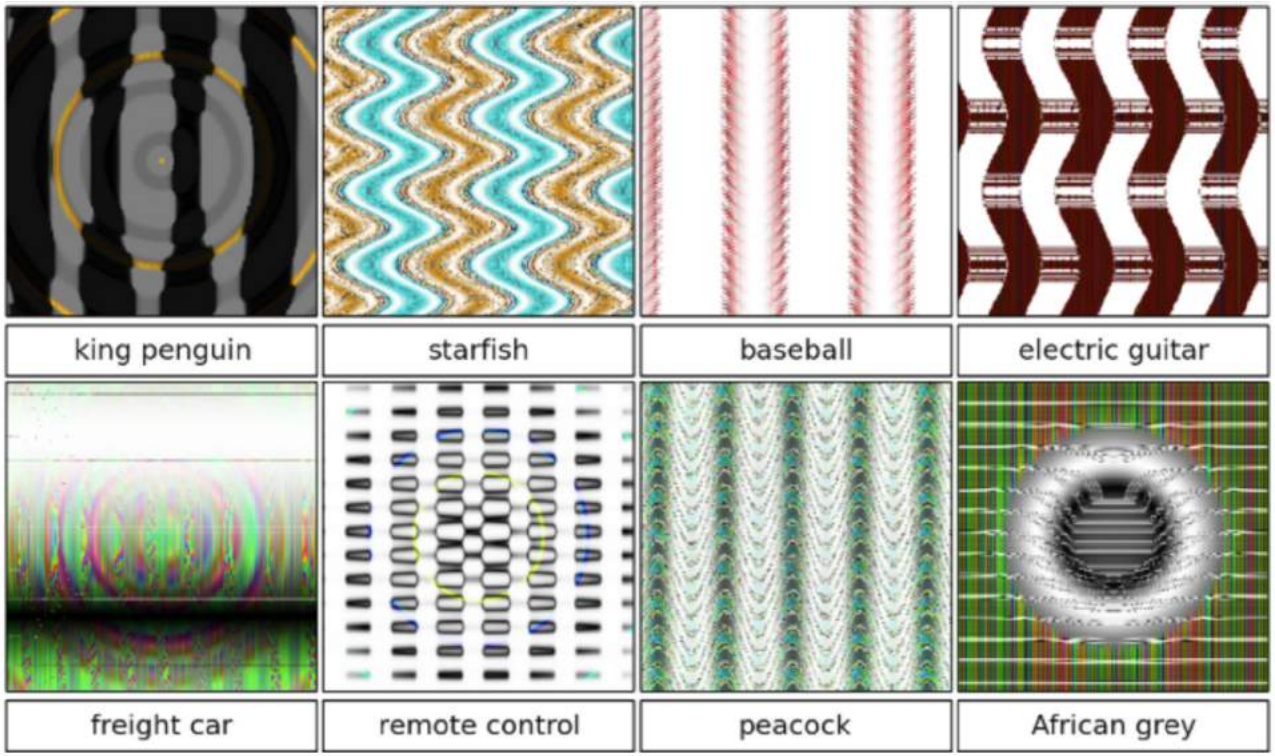
$=$

"panda"
57.7% confidence

"gibbon"
99.3 % confidence

Fooling ConvNets

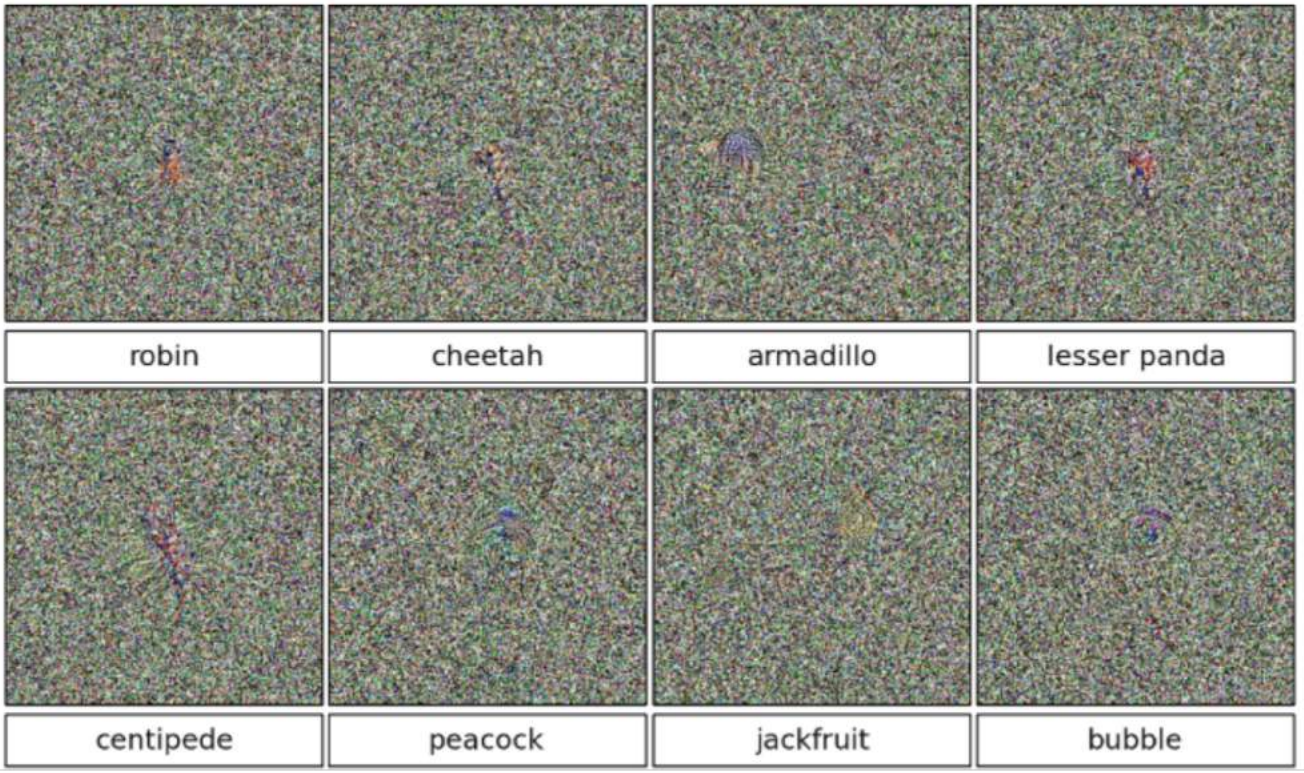
It is easy to generate perceptually meaningless images that will be classified as any given class with high confidence



A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, CVPR 2015

Fooling ConvNets

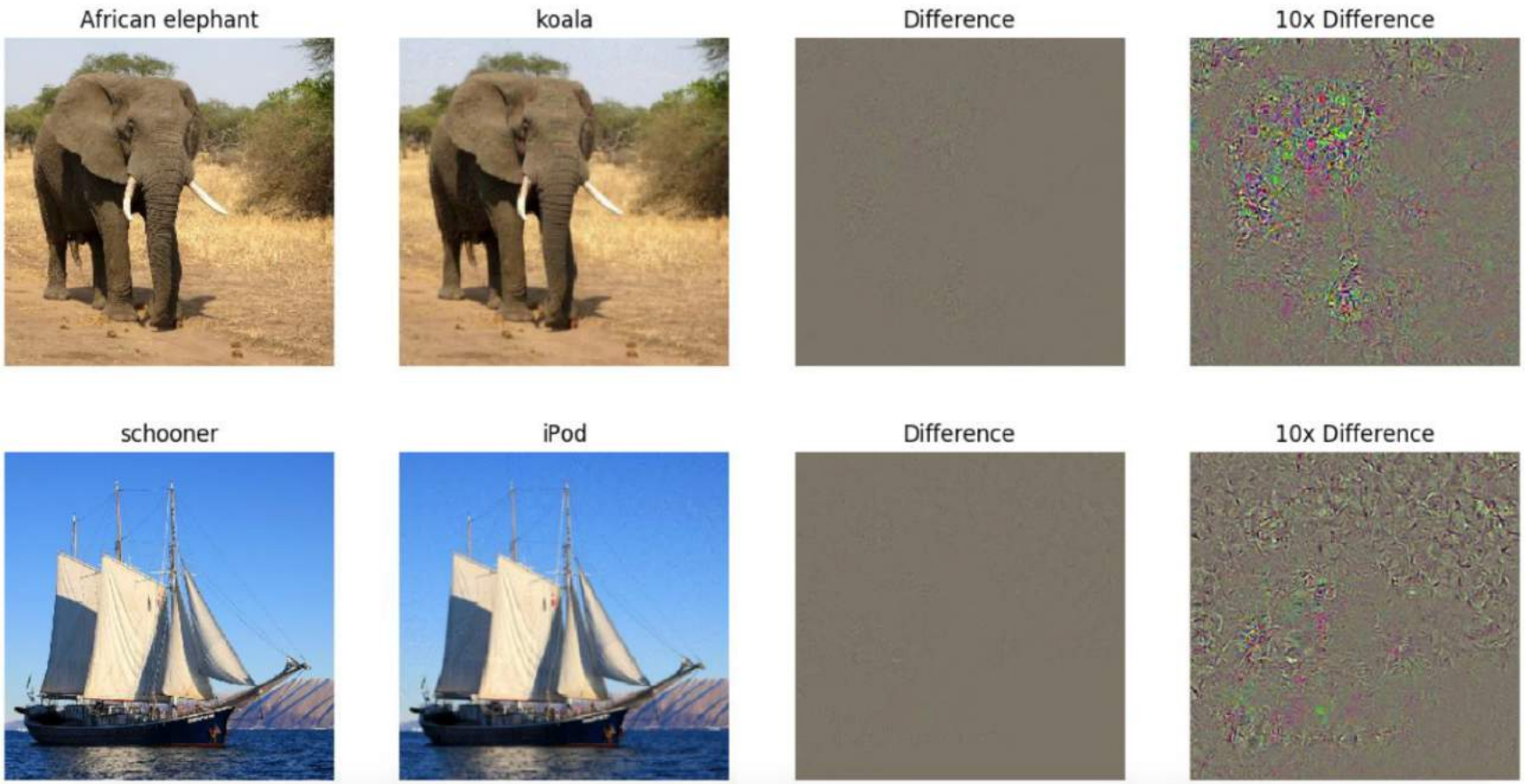
It is easy to generate perceptually meaningless images that will be classified as any given class with high confidence



A. Nguyen, J. Yosinski, J. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, CVPR 2015

Fooling ConvNets: Adversarial Samples

We can “fool” a neural network by imperceptibly perturbing an input image so it is misclassified



Source: Stanford CS231n

Example: Adversarial Samples for a Linear Classifier

- Increase score of target class $\hat{y}$:

$$x \leftarrow x + \eta \frac{\partial f(x, \hat{y})}{\partial x}$$

- For a linear classifier with $f(x, \hat{y}) = w^T x$:

$$x \leftarrow x + \eta\, w$$

- To fool a linear classifier, add a small multiple of the target class weights to the test example

Example: Adversarial Samples for a Linear Classifier

- Response of classifier with weights w to adversarial example $x + r$:

$$w^T(x + r) = w^T x + w^T r$$

- Suppose the classifier is normally expected to predict the same class for x and $x + r$ as long as $\|r\|_\infty \leq \epsilon$
- How to choose r to maximize the increase in activation $w^T r$ subject to $\|r\|_\infty \leq \epsilon$?

$$r = \epsilon \operatorname{sgn}(w)$$

I. Goodfellow, J. Schlens, C. Szegedy, Explaining and harnessing adversarial examples, ICLR 2015

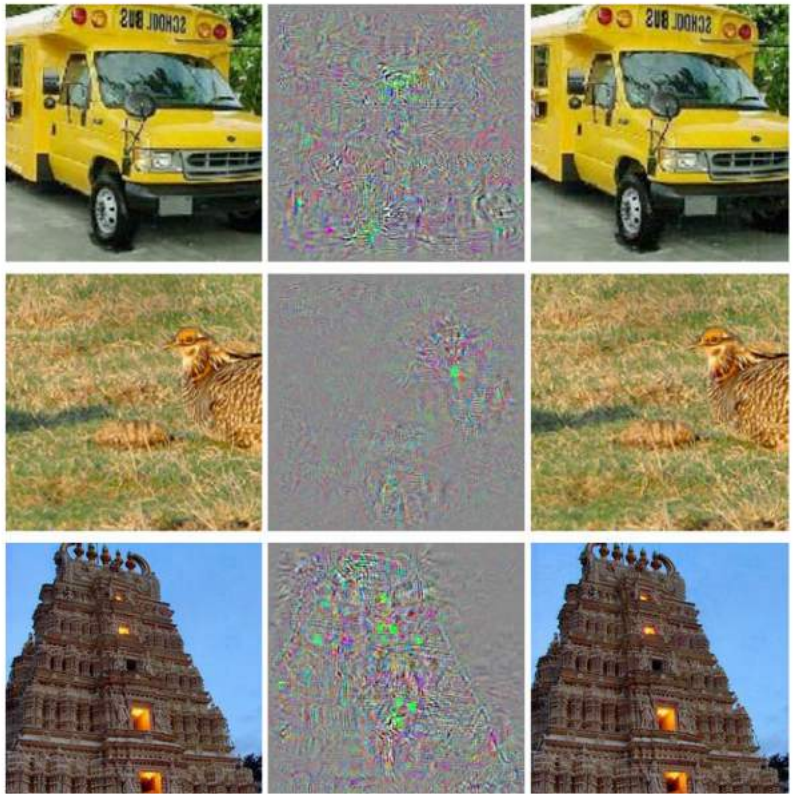
How to Find Adversarial Samples?

- Start with correctly classified image x
- Find perturbation r minimizing $\|r\|_2$ such that
  - $x + r$ is misclassified (or classified as specific target class)
  - All values of $x + r$ are in the valid range
- Constrained non-convex optimization, can be done using L-BFGS
- Solution: adversarial samples with smallest perturbation

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, ICLR 2014

How to Find Adversarial Samples?

Sample results:



Input | Perturbation x 10 | "Ostrich"

Input | Perturbation x 10 | "Ostrich"

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, ICLR 2014

How to Find Adversarial Samples?

- Rather than searching for the smallest possible perturbation, it is easier to take small gradient steps in desired direction (*remember the linear classifier case*?)
- Decrease score (increase loss) of correct class y^*:

$$x \leftarrow x + \eta \frac{\partial L(x, y^*)}{\partial x}$$

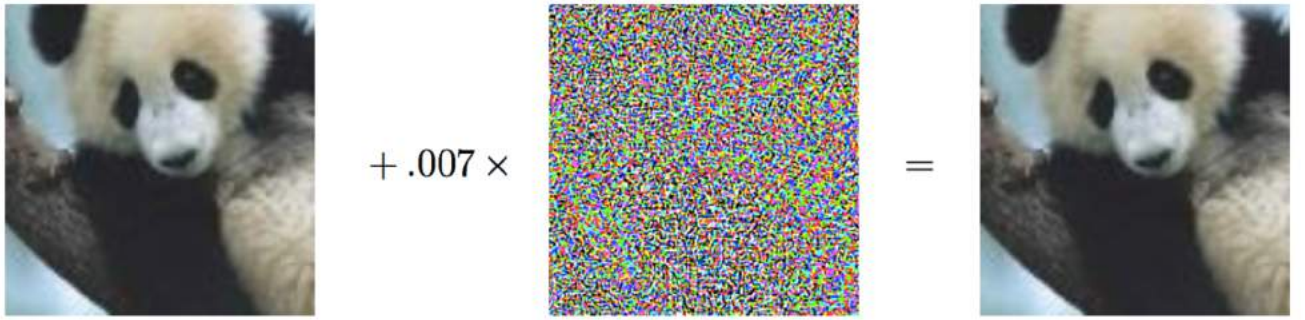
- Increase score (decrease loss) of incorrect target class $\hat{y}$:

$$x \leftarrow x - \eta \frac{\partial L(x, \hat{y})}{\partial x}$$

Adversarial Samples for ConvNets

Fast gradient sign method: Find the gradient of the loss w.r.t. correct class y^*, take element-wise sign, update in resulting direction:

$$x \leftarrow x + \epsilon \operatorname{sgn}\left(\frac{\partial L(x, y^*)}{\partial x}\right)$$



Goodfellow, J. Schlens, C. Szegedy,
Explaining and harnessing adversarial examples, ICLR 2015

Adversarial Samples for ConvNets

- **Fast gradient sign method:**

$$x \leftarrow x + \epsilon \, \mathrm{sgn}\left(\frac{\partial L(x, y^*)}{\partial x}\right)$$

- **Project gradient descent:** take multiple small steps, each time clip the result to be within ϵ-neighborhood of original image
- **Q:** How to choose the target label y^* ?

A. Kurakin, I. Goodfellow, S. Bengio,
Adversarial examples in the real world, ICLR 2017 workshop

Adversarial Samples for ConvNets

- **Fast gradient sign method:**

$$x \leftarrow x + \epsilon \operatorname{sgn}\left(\frac{\partial L(x, y^*)}{\partial x}\right)$$

- **Project gradient descent:** take multiple small steps, each time clip the result to be within ϵ-neighborhood of original image
- ***Least likely class* method:** try to misclassify image as class $\hat{y}$ with smallest initial score:

$$x \leftarrow x - \epsilon \operatorname{sgn}\left(\frac{\partial L(x, \hat{y})}{\partial x}\right)$$

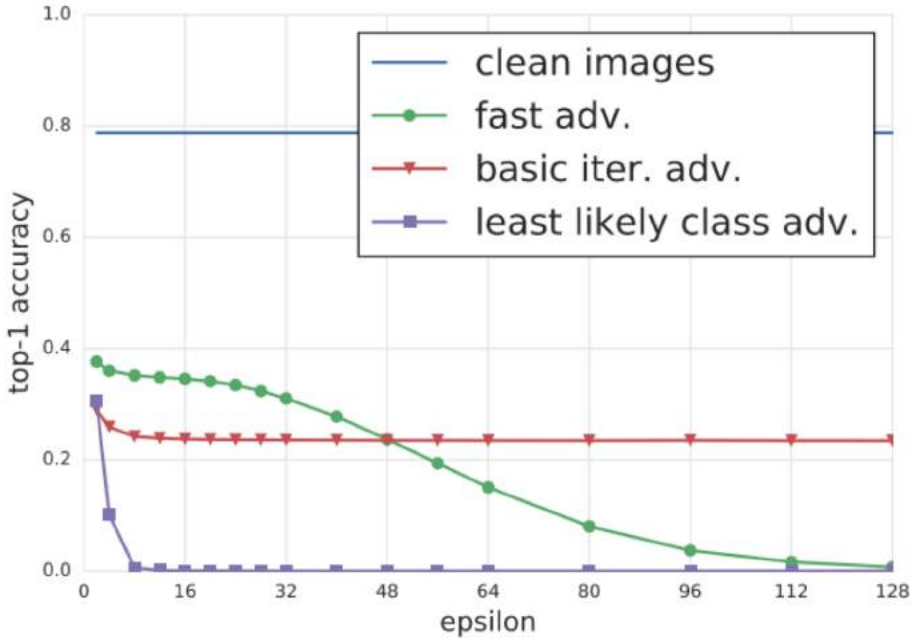
A. Kurakin, I. Goodfellow, S. Bengio,
Adversarial examples in the real world, ICLR 2017 workshop

Adversarial Samples for ConvNets

Comparison of methods for $\epsilon = 32$

A. Kurakin, I. Goodfellow, S. Bengio,
Adversarial examples in the real world, ICLR 2017 workshop

Adversarial Samples for ConvNets



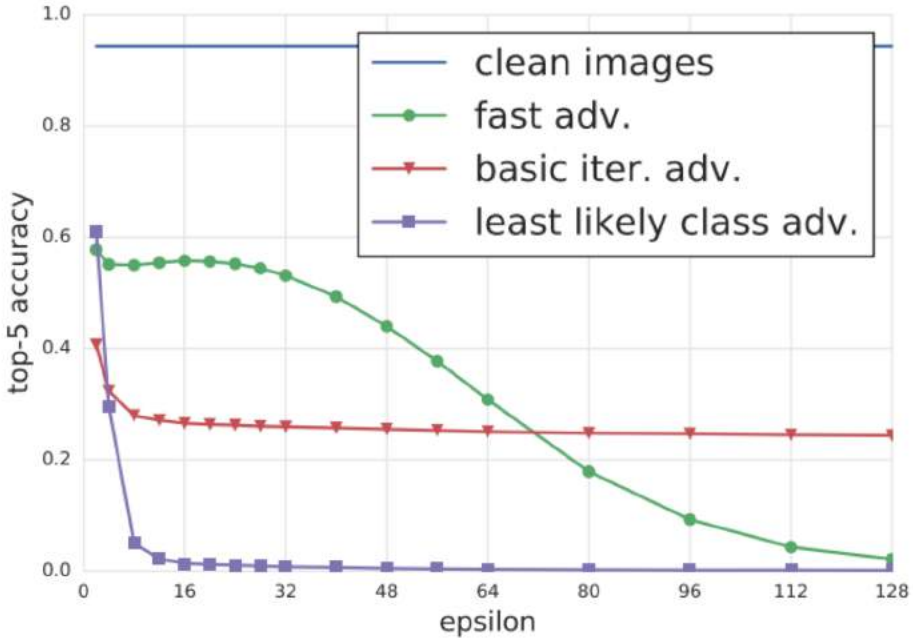


Figure 2: Top-1 and top-5 accuracy of Inception v3 under attack by different adversarial methods and different ϵ compared to "clean images" — unmodified images from the dataset. The accuracy was computed on all $50,000$ validation images from the ImageNet dataset. In these experiments ϵ varies from 2 to 128.

A. Kurakin, I. Goodfellow, S. Bengio,
Adversarial examples in the real world, ICLR 2017 workshop

Adversarial Samples vs. Interpretability

- **Fast gradient sign method:**

$$x \leftarrow x + \epsilon \operatorname{sgn}\left(\frac{\partial L(x, y^*)}{\partial x}\right)$$

- **Project gradient descent:** take multiple small steps, each time clip the result to be within ϵ-neighborhood of original image
- ***Least likely class* method:** try to misclassify image as class $\hat{y}$ with smallest initial score:

$$x \leftarrow x - \epsilon \operatorname{sgn}\left(\frac{\partial L(x, \hat{y})}{\partial x}\right)$$

A. Kurakin, I. Goodfellow, S. Bengio,
Adversarial examples in the real world, ICLR 2017 workshop

Adversarial Samples vs. Interpretability

- **Fast gradient sign method:**

$$x \leftarrow x + \epsilon \operatorname{sgn}\left(\frac{\partial L(x, y^*)}{\partial x}\right)$$

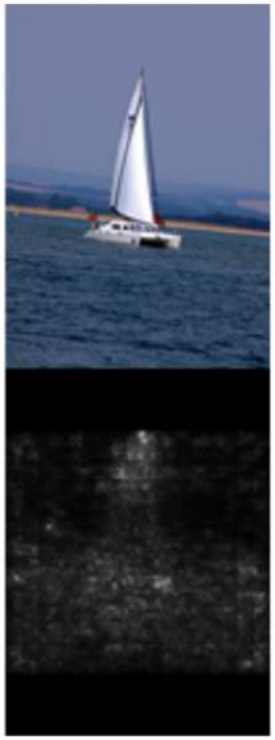
- **Project gradient descent:** take multiple small steps, each time clip the result to be within ϵ-neighborhood of original image
- ***Most likely class* method:** try to identify pixels that support the most confident $\hat{y}$

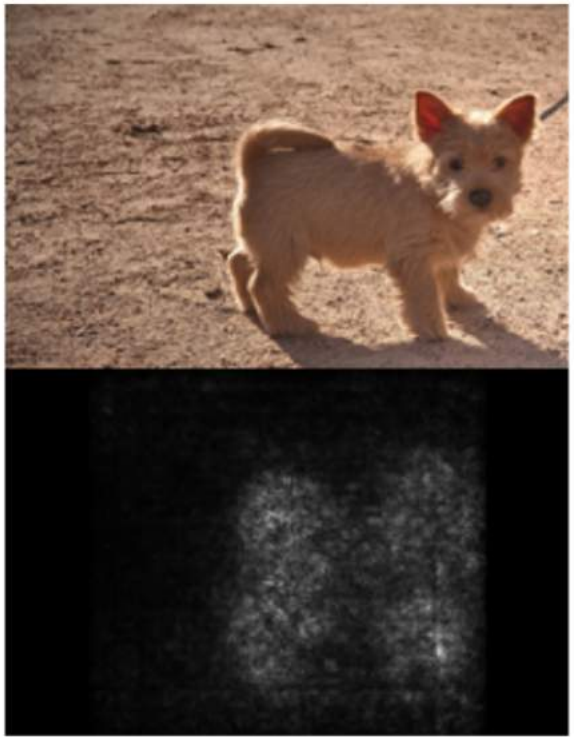
$$x \leftarrow x - \epsilon \operatorname{sgn}\left(\frac{\partial L(x, \hat{y})}{\partial x}\right)$$

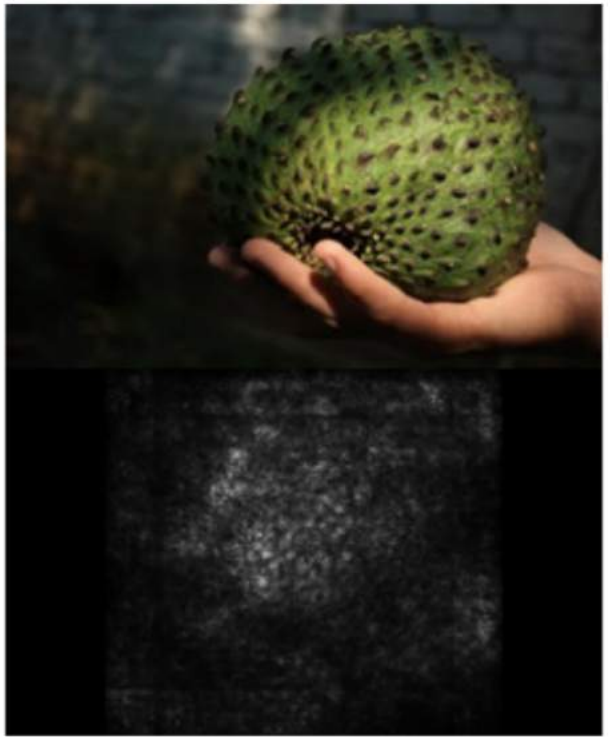
A. Kurakin, I. Goodfellow, S. Bengio,
Adversarial examples in the real world, ICLR 2017 workshop

“White Box” Saliency: Using Gradients

Compute $-\frac{\partial L(x;\theta)}{x}$ and display the max of absolution values across three channels



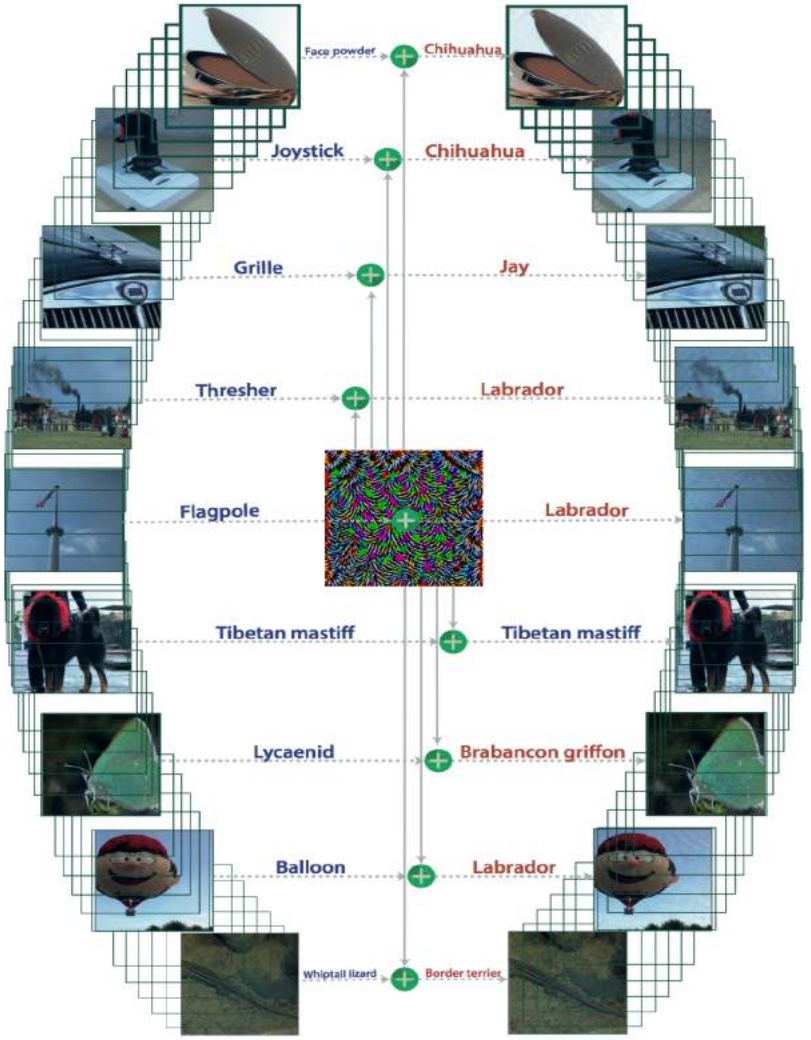




K. Simonyan, A. Vedaldi, and A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, ICLR 2014

Universal Adversarial Perturbations

- Goal: for a given network, find an *image-independent* perturbation vector that causes *all images* to be misclassified with high probability



S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, CVPR 2017

Universal Adversarial Perturbations

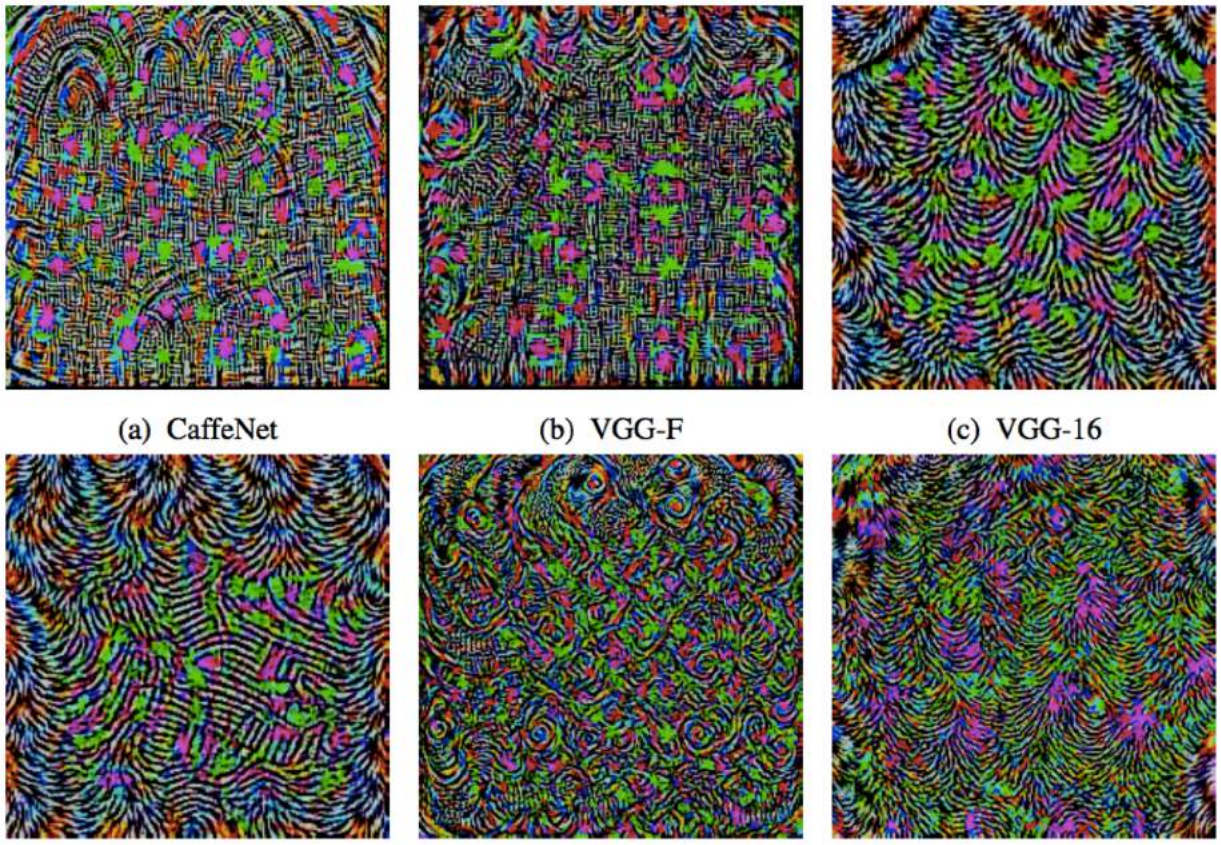
Approach:

- Start with $r = 0$
- Cycle through training examples x_i (in multiple passes)
  - If $x_i + r$ is misclassified, skip to x_{i+1}
  - Find minimum perturbation Δr that takes $x_i + r + \Delta r$ to another class
  - Update $r \leftarrow r + \Delta r$, enforce $\|r\| \leq \epsilon$
- Terminate when fooling rate on training examples reaches target value

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, CVPR 2017

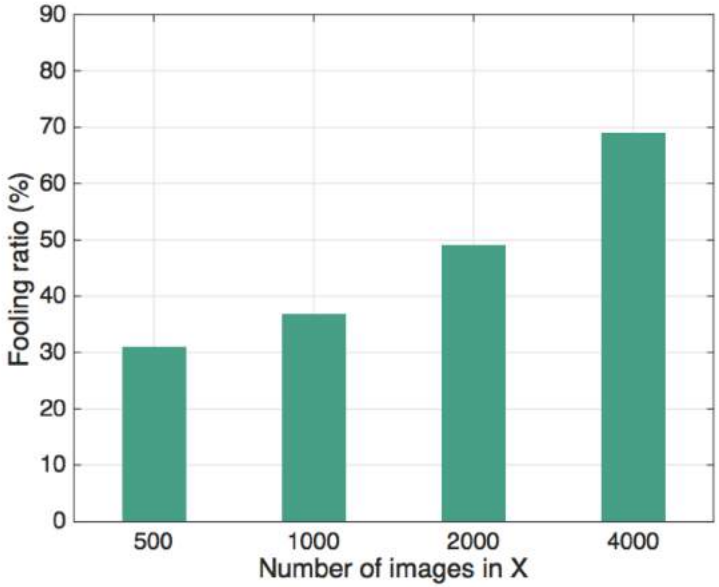
Universal Adversarial Perturbations

Perturbation vectors computed from different architectures:



S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard, Universal adversarial perturbations, CVPR 2017

Universal Adversarial Perturbations



Fooling ratio on validation set vs. training set size for GoogLeNet

Fooling rates on different models after training on 10,000 images

		CaffeNet [8]	VGG-F [2]	VGG-16 [17]	VGG-19 [17]	GoogLeNet [18]	ResNet-152 [6]
ℓ_2	X	85.4%	85.9%	90.7%	86.9%	82.9%	89.7%
	Val.	85.6	87.0%	90.3%	84.5%	82.0%	88.5%
ℓ_∞	X	93.1%	93.8%	78.5%	77.8%	80.8%	85.4%
	Val.	93.3%	93.7%	78.3%	77.8%	78.9%	84.0%

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard,
Universal adversarial perturbations, CVPR 2017

Universal Adversarial Perturbations

Universal perturbations turn out to generalize well across models!

Fooling rate when computing a perturbation for one model (rows) and testing it on others (columns)

	VGG-F	CaffeNet	GoogLeNet	VGG-16	VGG-19	ResNet-152
VGG-F	**93.7%**	71.8%	48.4%	42.1%	42.1%	47.4 %
CaffeNet	74.0%	**93.3%**	47.7%	39.9%	39.9%	48.0%
GoogLeNet	46.2%	43.8%	**78.9%**	39.2%	39.8%	45.5%
VGG-16	63.4%	55.8%	56.5%	**78.3%**	73.1%	63.4%
VGG-19	64.0%	57.2%	53.6%	73.5%	**77.8%**	58.0%
ResNet-152	46.3%	46.3%	50.5%	47.0%	45.5%	**84.0%**

S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, P. Frossard,
Universal adversarial perturbations, CVPR 2017

Black Box Adversarial Attacks

- Suppose the adversary can only query a target network with chosen inputs and observe the outputs
- Key idea: learn substitute for target network using synthetic input data, use substitute network to craft adversarial examples
- Successfully attacked third-party APIs from MetaMind, Amazon, and Google, but only on low-res digit and street sign images

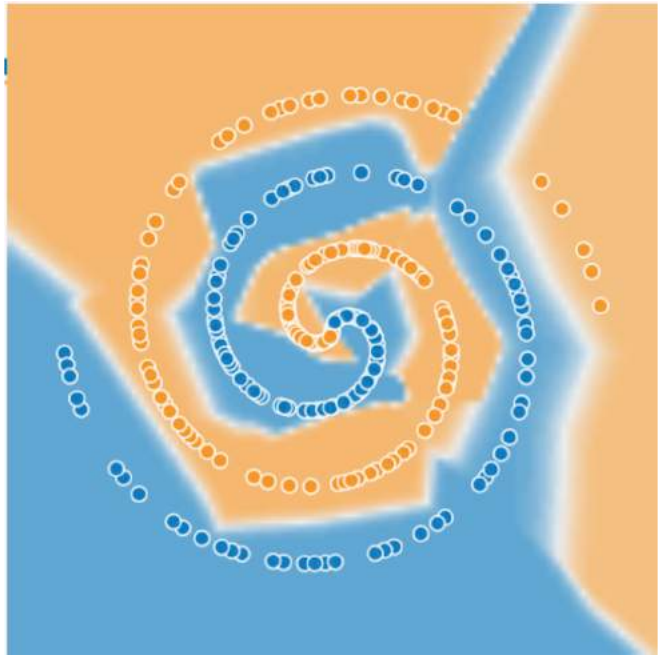
N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical Black-Box Attacks against Machine Learning, ACM Asia Conference on Computer and Communications Security, 2017

Adversarial Samples: Recap

- For any input image, it is usually easy to generate a very similar image that gets misclassified by the same network
- To obtain an adversarial example, one does not need to do precise gradient ascent
- It is possible to attack many images with the same perturbation
- Adversarial examples that can fool one network have a high chance of fooling a network with different parameters and even architecture

Adversarial Samples: Why???

- Counter-intuitively, a network can both generalize well on natural images and be susceptible to adversarial examples
- Networks (with Conv+ReLU) are piece-wise linear
  - Too many pieces in high dimensional space
  - ”Easy” to find a way that can cross the decision boundary
  - Still… why universal patterns across models?



https://playground.tensorflow.org

G.H Lee, D. Alvarez-Melis, T. S. Jaakkola, Towards Robust, Locally Linear Deep Networks, ICLR'18

Defending against Adversarial Samples

Adversarial training: networks can be made somewhat resistant by augmenting or regularizing training with adversarial examples

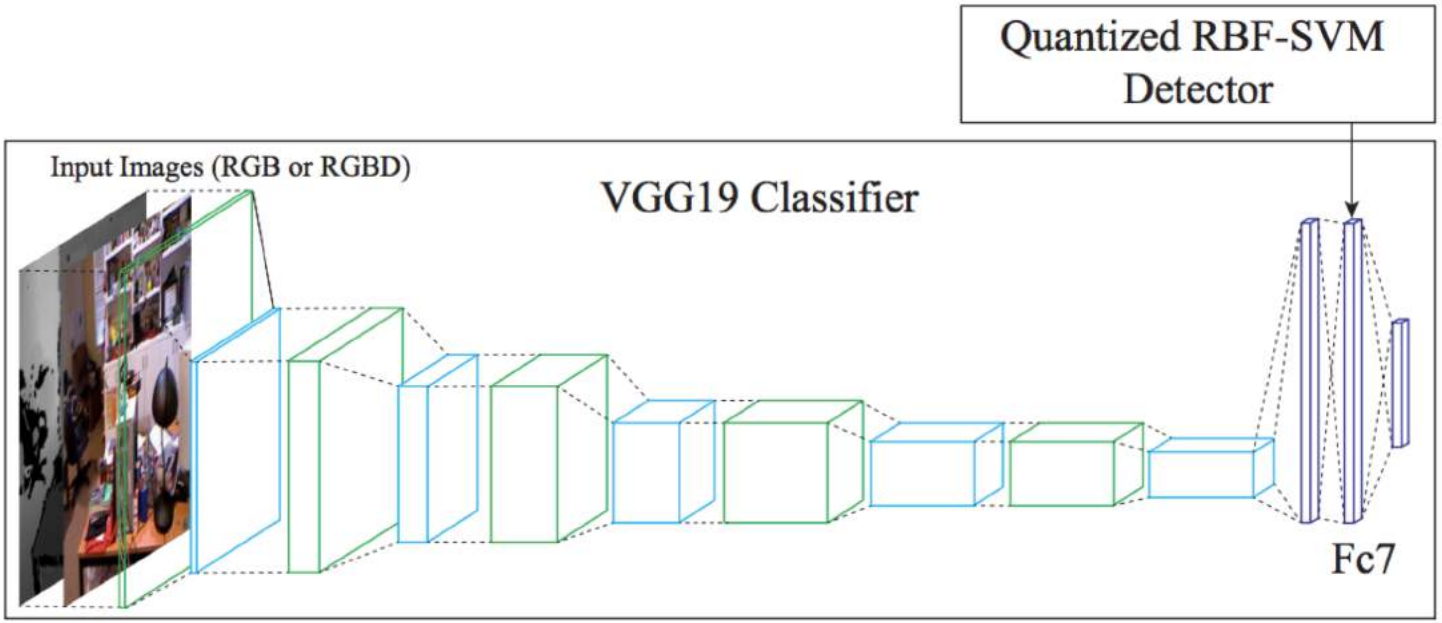
$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{S}} L(\theta, x+\delta, y) \right]$$

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu,
Towards Deep Learning Models Resistant to Adversarial Attacks, ICLR 2018

F. Tramer, A. Kurakin, N. Papernot, D. Boneh, P. McDaniel,
Ensemble adversarial training: Attacks and defenses, ICLR 2018

Defending against Adversarial Samples

Train a separate model to reject adversarial examples: SafetyNet



J. Lu, T. Issaranon, D. Forsyth,
SafetyNet: Detecting and Rejecting Adversarial Examples Robustly, CVPR 2017

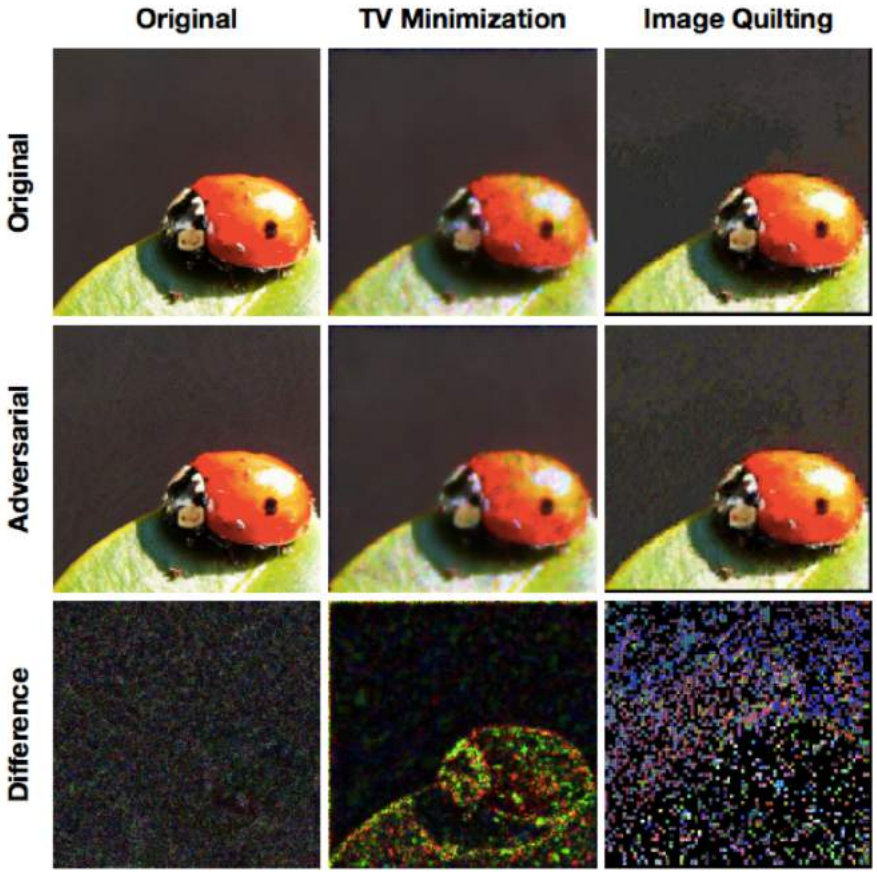
Defending against Adversarial Samples

Design highly nonlinear architectures robust to adversarial perturbations

D. Krotov, J. Hopfield,
Dense Associative Memory is Robust to Adversarial Inputs, arXiv 2017

Defending against Adversarial Samples

Pre-process input images to disrupt adversarial perturbations

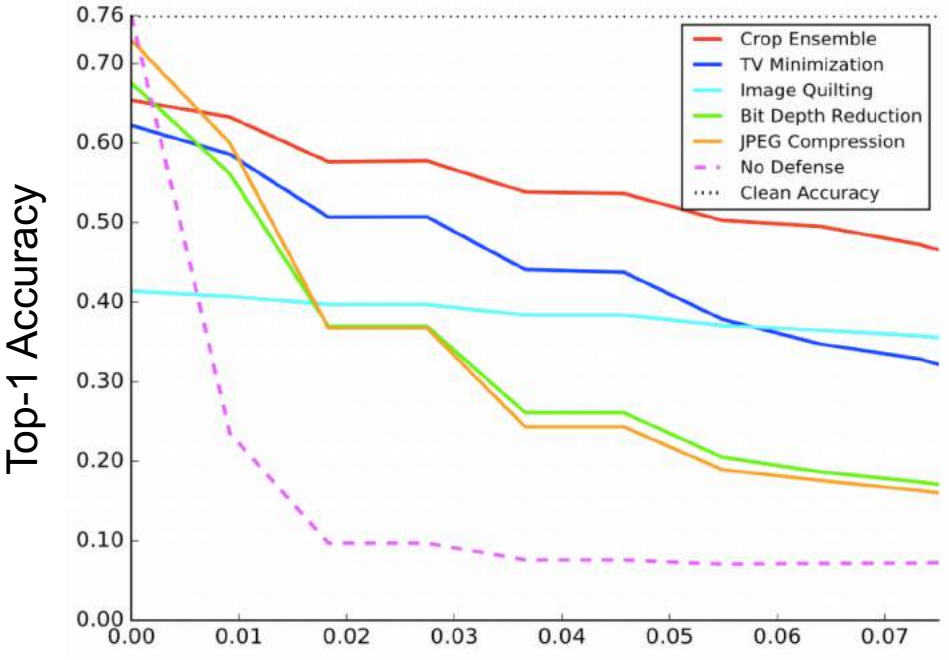


C. Guo, M. Rana, M. Cisse, L. van der Maaten, Countering Adversarial Images Using Input Transformations, ICLR 2018

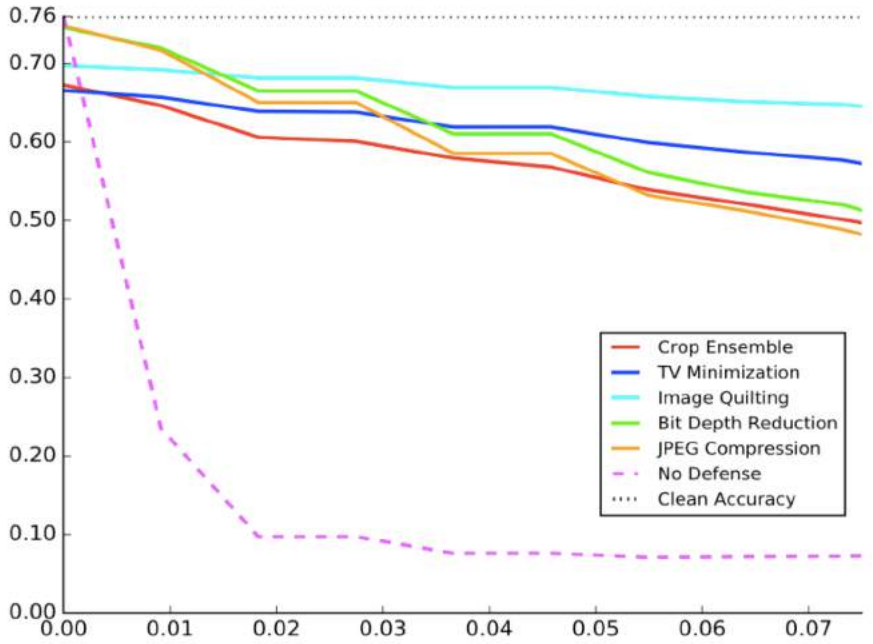
Defending against Adversarial Samples

Pre-process input images to disrupt adversarial perturbations

ResNet-50 *tested* on transformed images



ResNet-50 *trained and tested* on transformed images



C. Guo, M. Rana, M. Cisse, L. van der Maaten, Countering Adversarial Images Using Input Transformations, ICLR 2018

Adversarial Samples & Human Vision

Adversarial examples that are designed to transfer across multiple architectures can also be shown to confuse the human visual system in rapid presentation settings

G. Elsayed et al., Adversarial Examples that Fool both Computer Vision and Time-Limited Humans, arXiv 2018