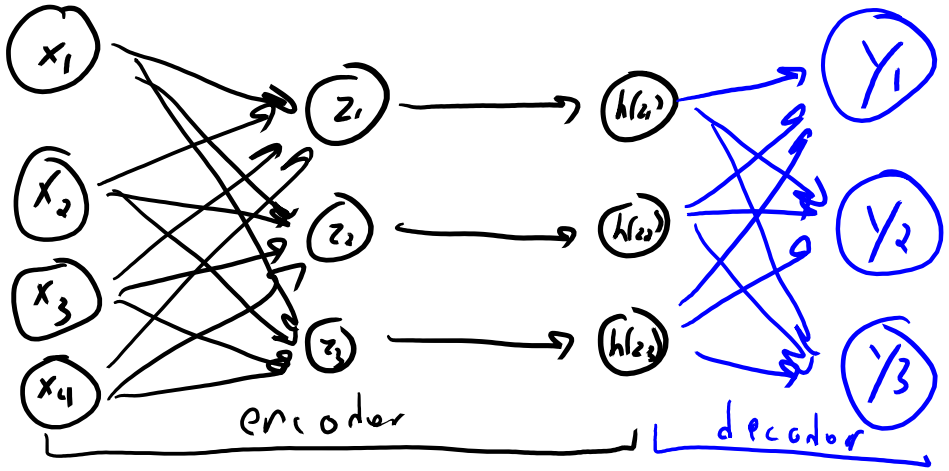
CPSC 440: Machine Learning

What do we learn?

Winter 2022

Last Time: Multi-Class Neural Networks

- We discussed multi-class classification with neural networks:



$$\hat{y}_1 = v_1^T h(Wx)$$

$$\hat{y}_2 = v_2^T h(Wx)$$

$$\hat{y}_3 = v_3^T h(Wx)$$

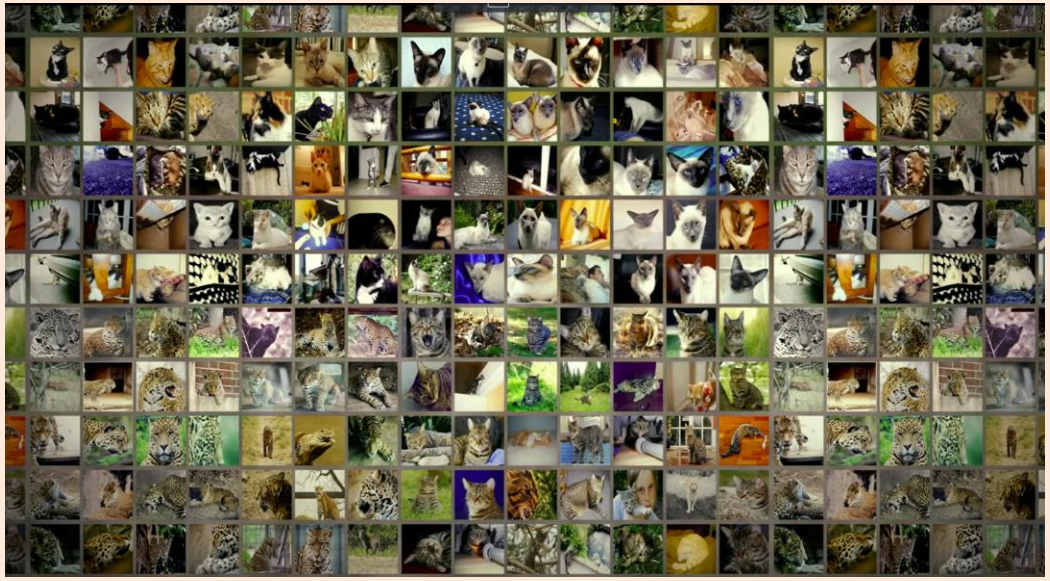
- We use the softmax function to convert the $\hat{y}_c$ to probabilities:
  - We use this for inference.
  - Likelihood is softmax for true label.
  - Last layer is all that changes.

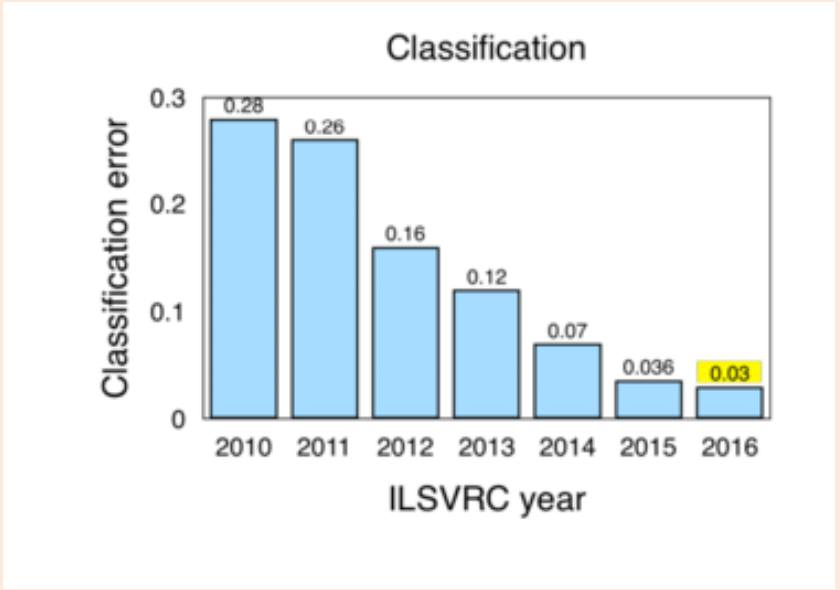
$$p(y = c \mid x, W, V) = \frac{\exp(\hat{y}_c)}{\sum_{c'=1}^{K} \exp(\hat{y}_{c'})}$$

- We train by minimizing the sum of negative log-likelihoods over 'i'.
  - We can add multiple layers, convolution layers, max pooling, ReLu, and so on.

Previously: ImageNet Competition and CNNs

- ImageNet: Millions of labeled images, 1000 object classes.
  - Led to popularization of CNNs and deep learning across computer vision.
  - Led to many insights about how to train CNNs and construct architectures.
    - ImageNet + CNNs is arguably most influential computer vision work of all time.

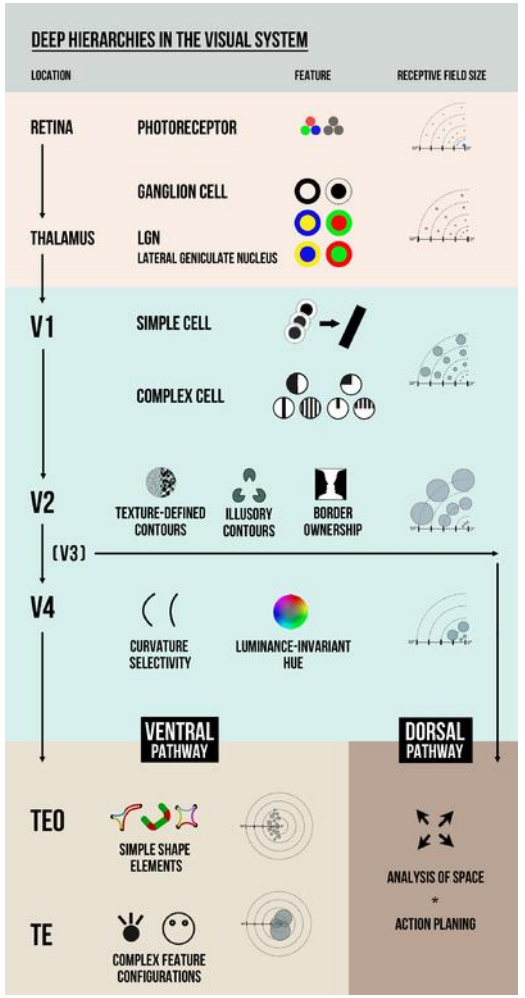


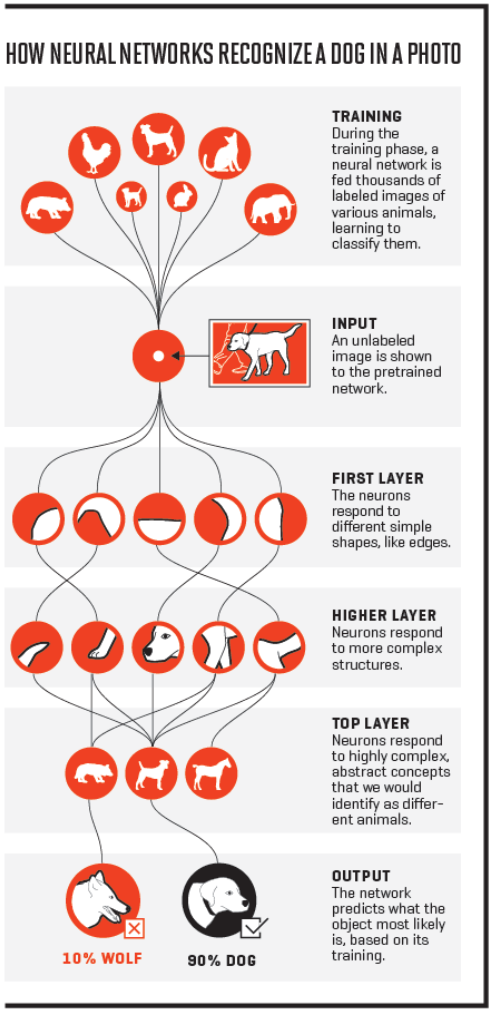


https://www.youtube.com/watch?v=40riCqvRoMs
http://www.themtank.org/a-year-in-computer-vision

Are CNNs learning something sensible?

- Recall that deep learning and CNNs are motivated by ideas about human vision.
  - First layers detect simple features like Gaussians, Gabors, Laplacian of Gaussian.
  - Later layers detect more complicated features like corners, repeating patterns.
  - Deeper layers starts to recognize complex parts of objects.
  - Deepest layers recognize full object concepts.
- Is this what trained CNNs actually do?





https://en.wikibooks.org/wiki/Sensory_Systems/Visual_Signal_Processing
http://fortune.com/ai-artificial-intelligence-deep-machine-learning/

Are CNNs learning something sensible?

- Filters learned by first layer of original AlexNet (first CNN winner):

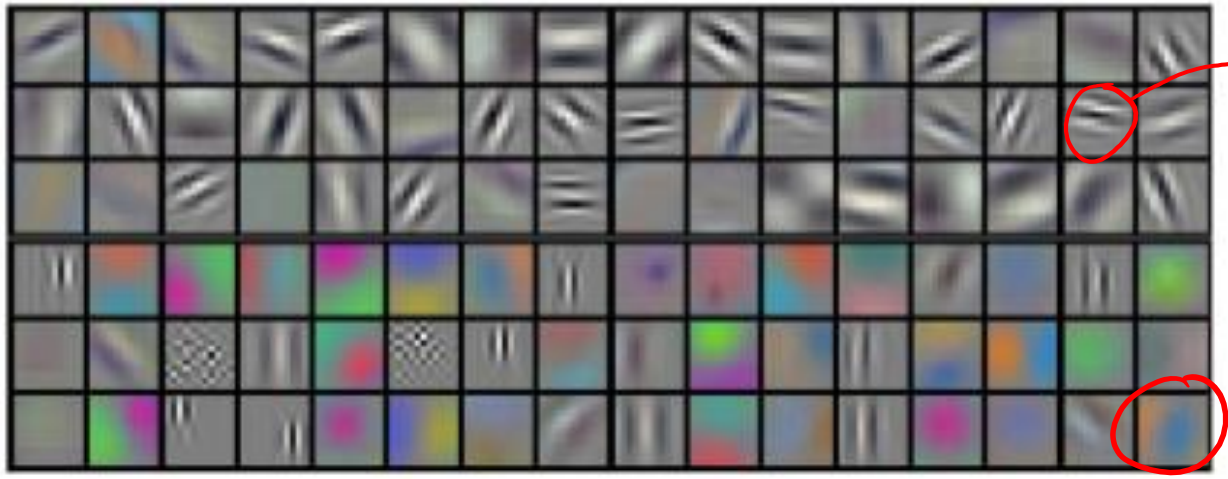


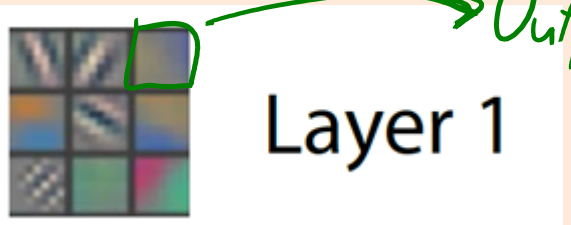
Figure 3: 96 convolutional kernels of size $11 \times 11 \times 3$ learned by the first convolutional layer on the $224 \times 224 \times 3$ input images. The

- Many other models give similar results (but often only 1 layer).

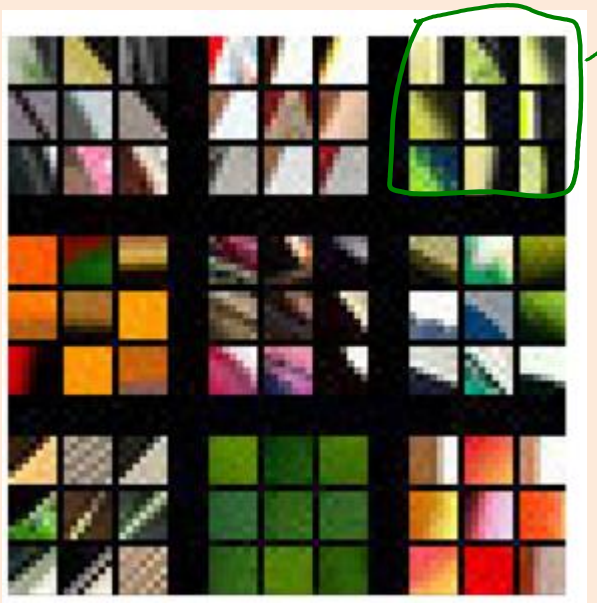
http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Are CNNs learning something sensible?

- It's harder to visualize what is learned in other layers.
  - Approach 1:
    - Search for training data image patches that maximally-activates a filter.
    - Then try to reason about what the filter is doing.
  - Approach 2:
    - Apply deconvolution network to these patches to try to "reverse" the operations.
    - Uses transposed convolutions and unpooling to visualize "what activated the filter".



Output of deconvolution network



Real images where filter responds most strongly.

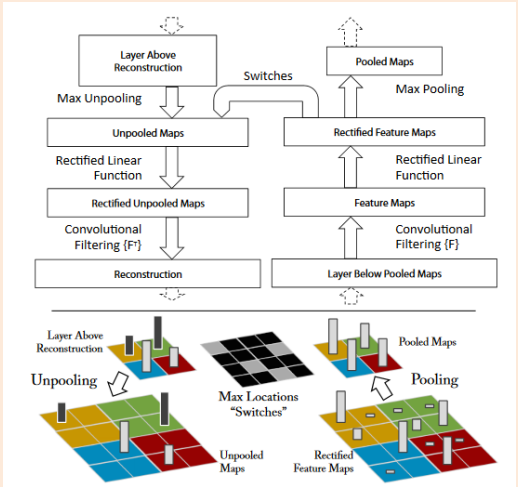
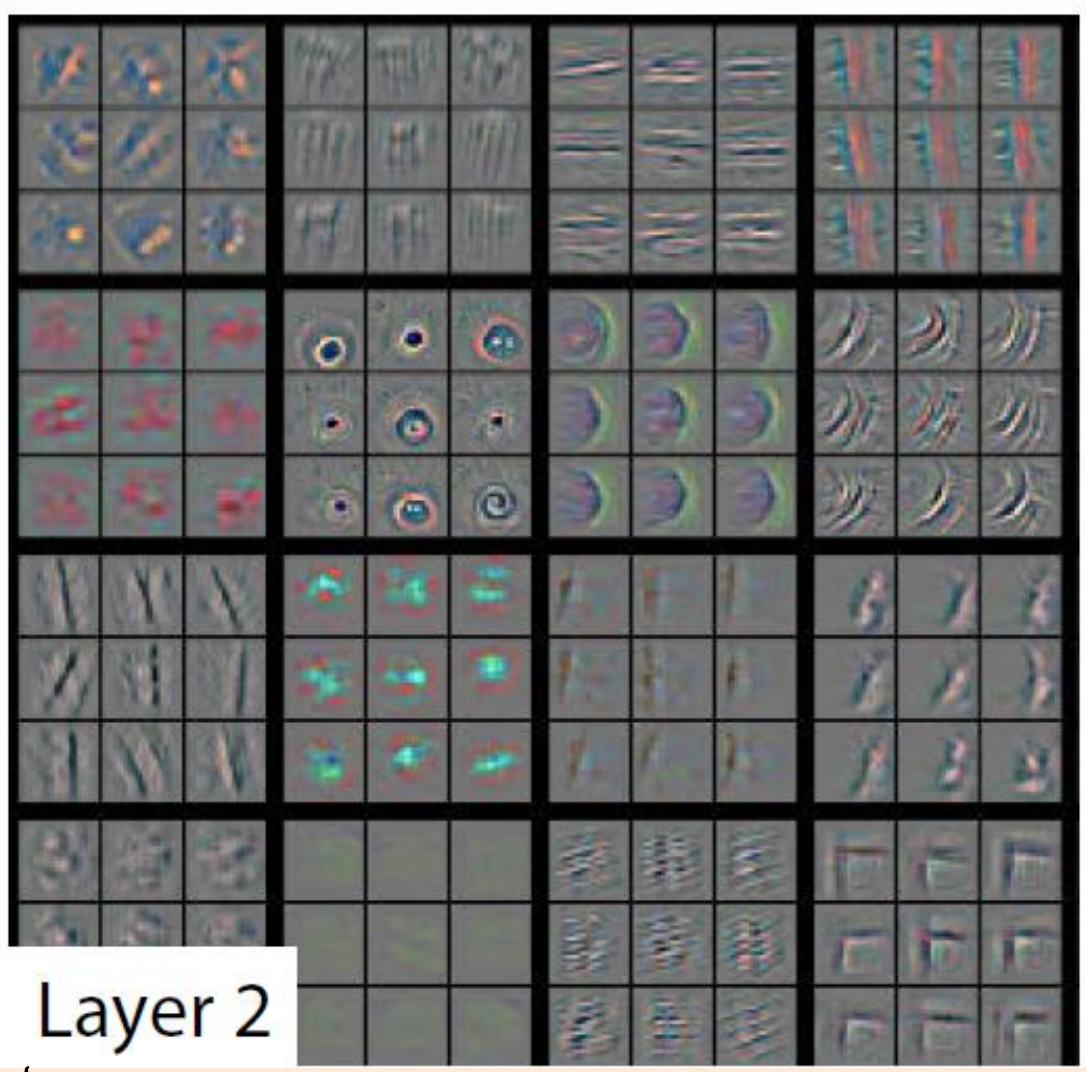
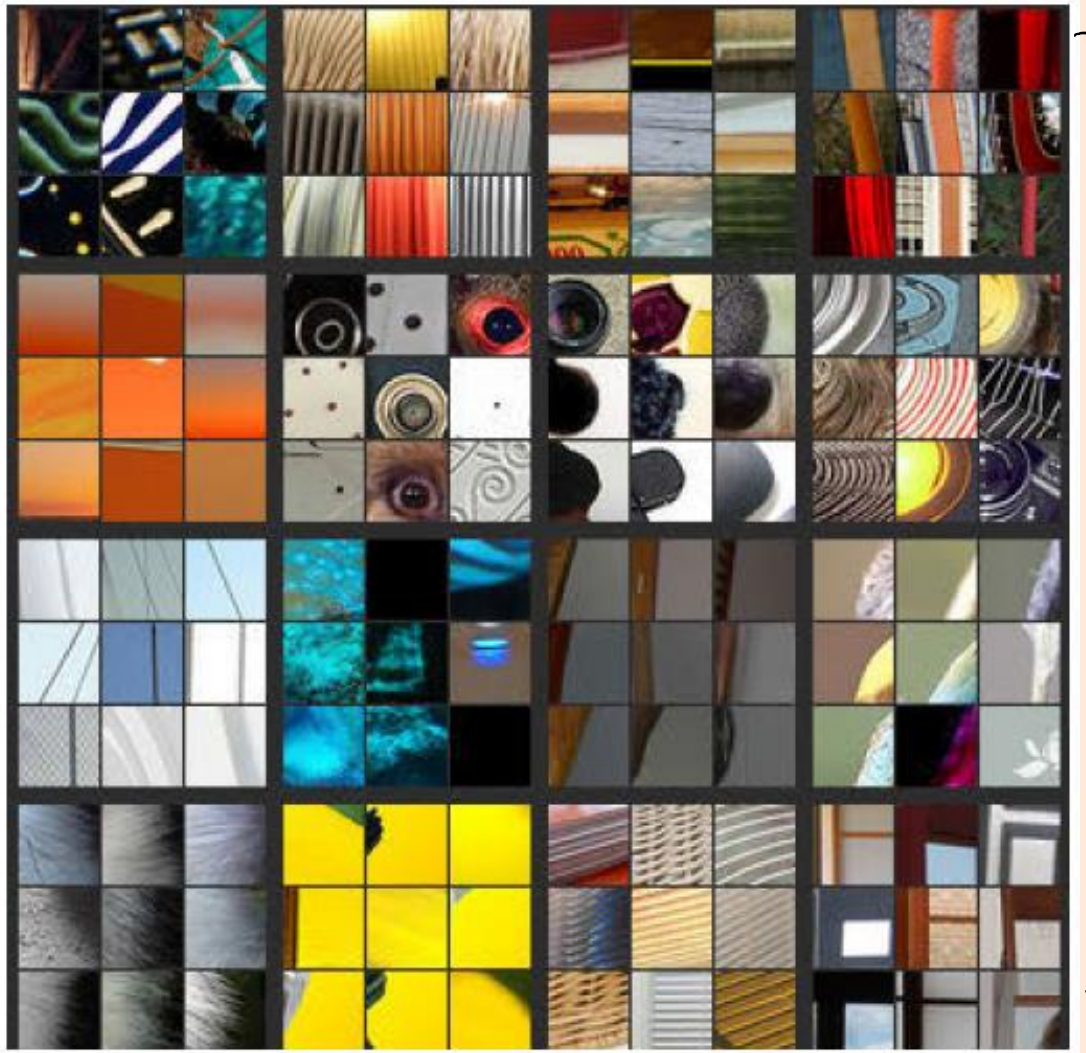


Figure 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local max in each pooling region (colored zones) during pooling in the convnet.

https://arxiv.org/pdf/1311.2901v3.pdf

Are CNNs learning something sensible?



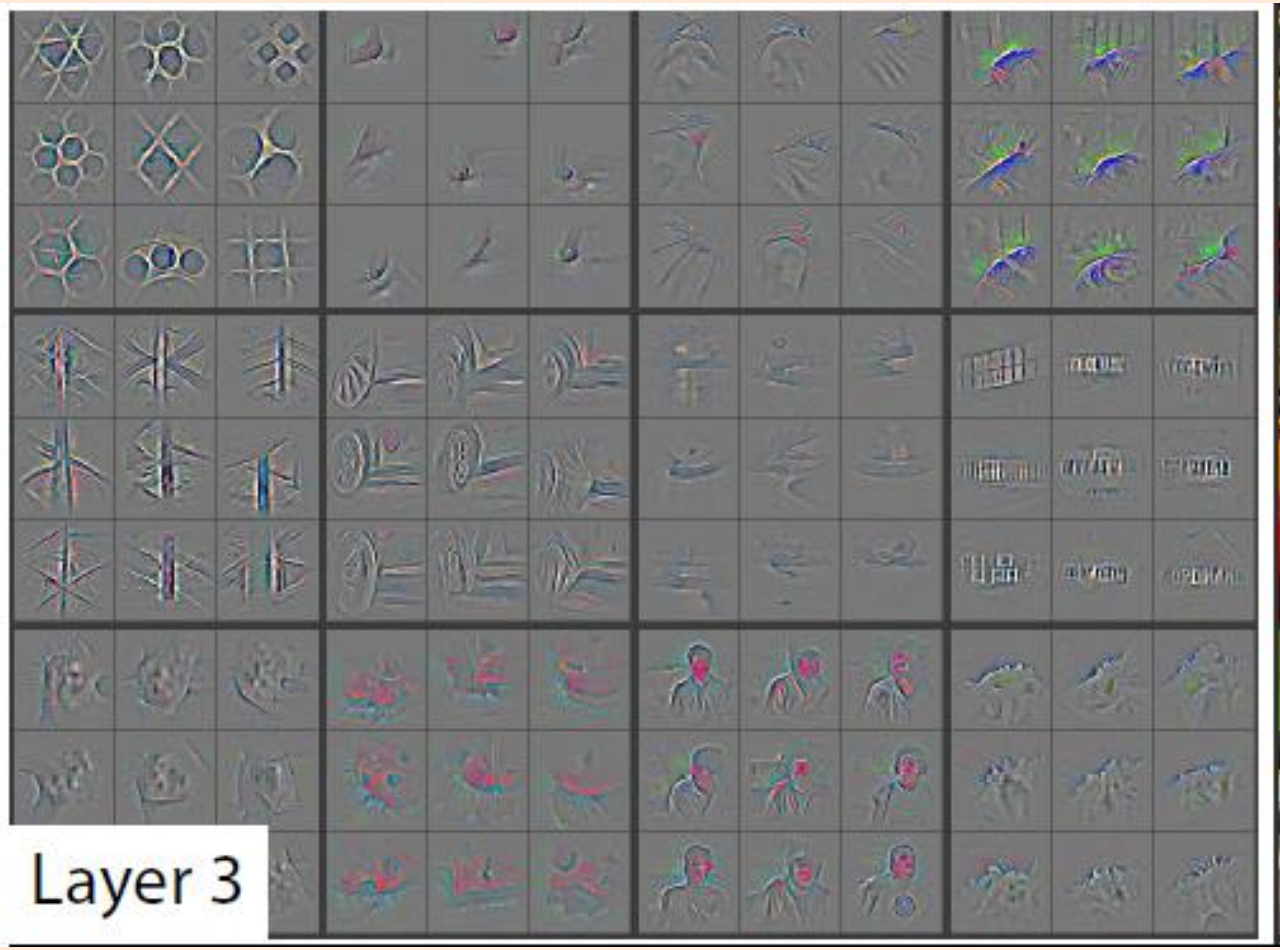


Patch from data giving largest response

Result of deconvolution network

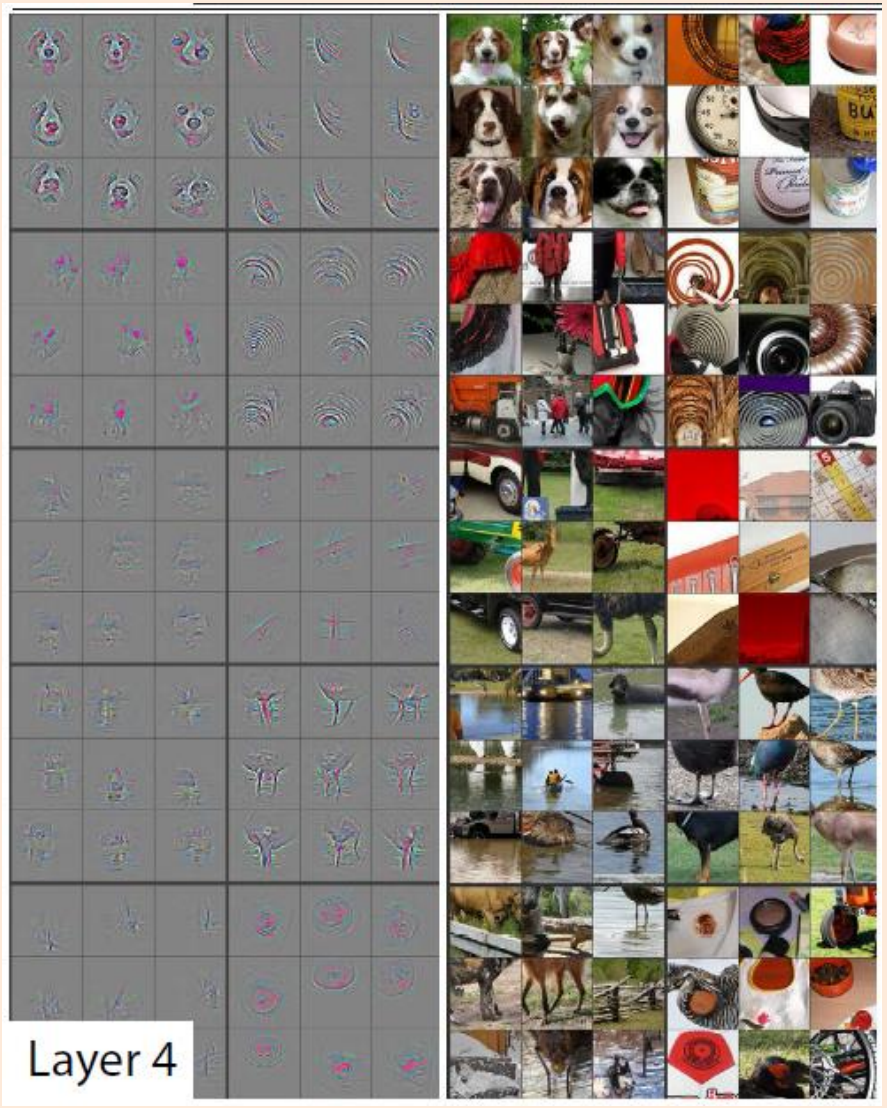
https://arxiv.org/pdf/1311.2901v3.pdf

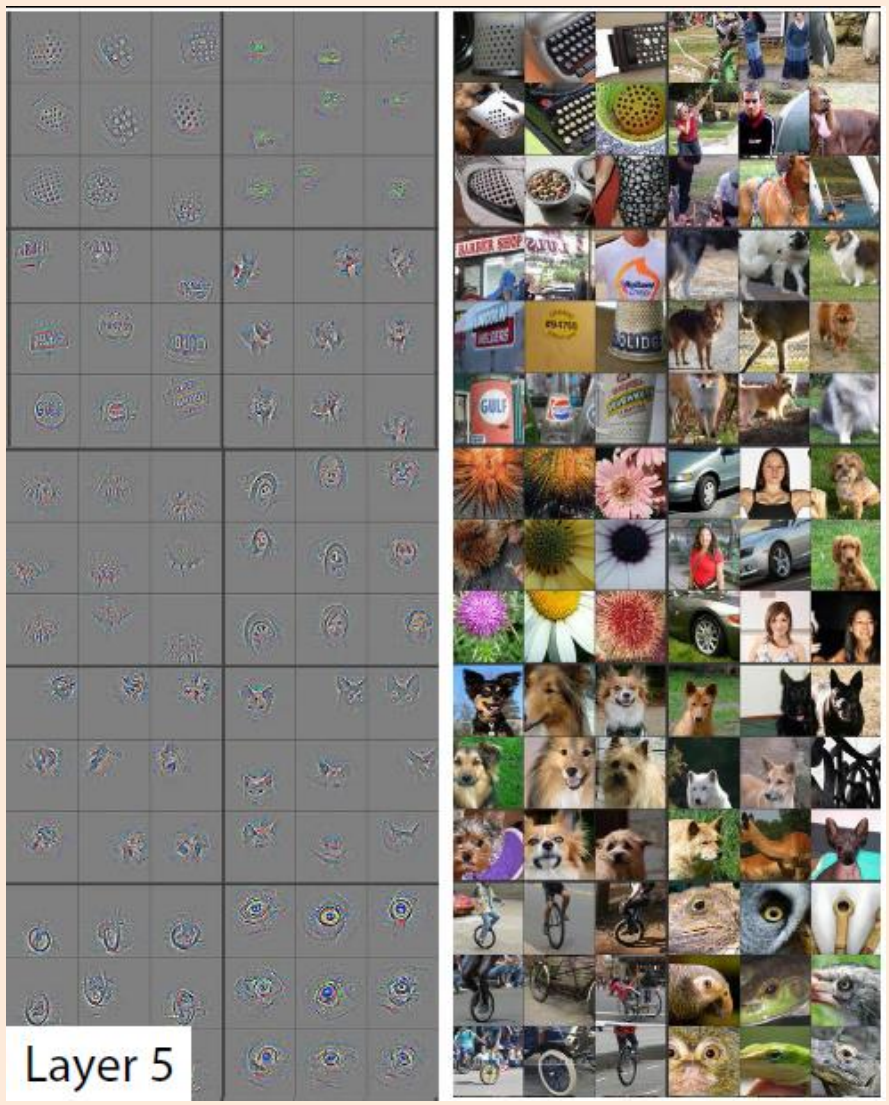
Are CNNs learning something sensible?



https://arxiv.org/pdf/1311.2901v3.pdf

Are CNNs learning something sensible?

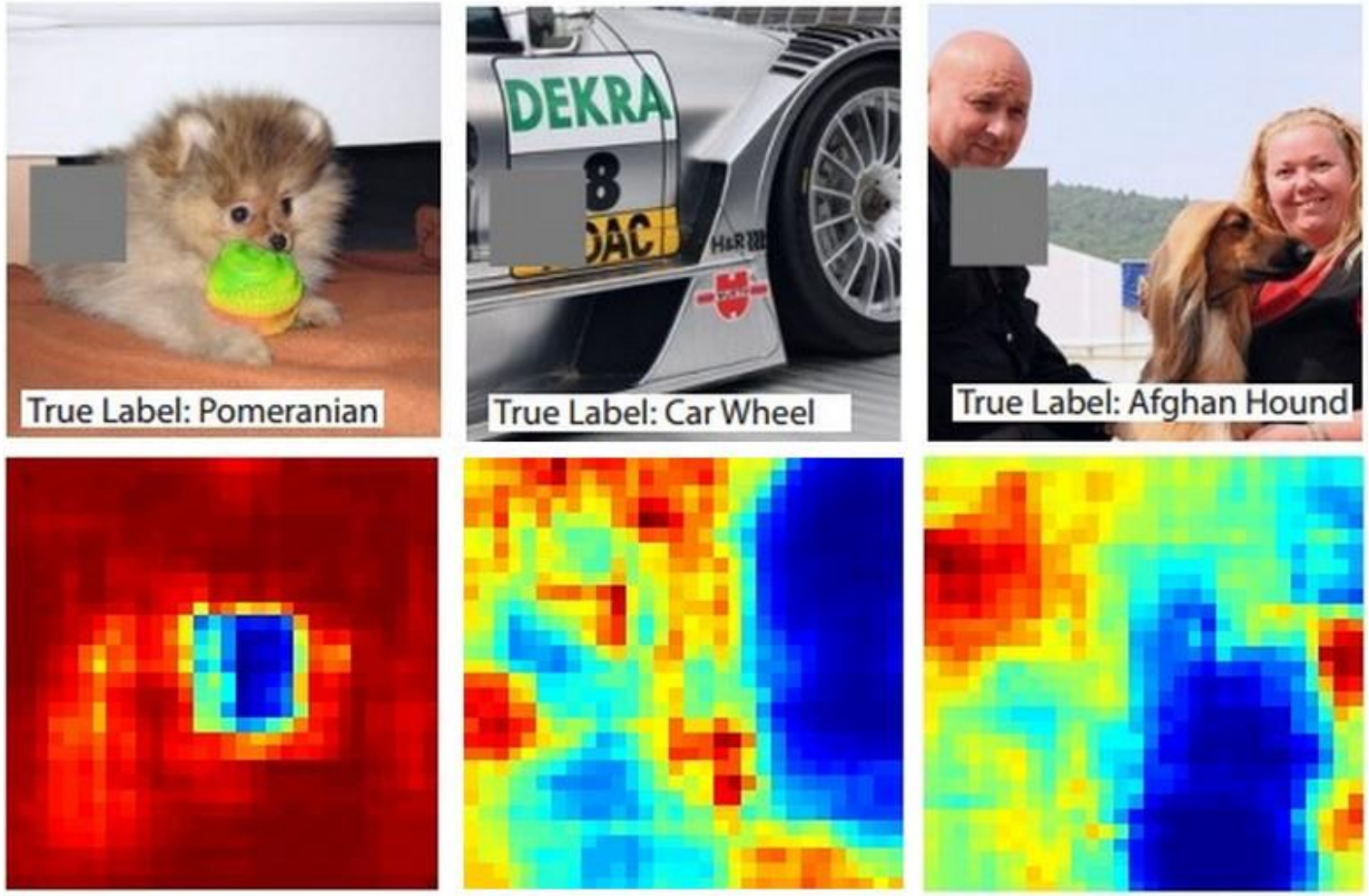




https://arxiv.org/pdf/1311.2901v3.pdf

Are CNNs learning something sensible?

- We can look at how prediction changes if we hide part of image:



Mission Accomplished?

- For speech recognition and object detection:
  - No non-deep methods have ever given the current level of performance.
  - Deep models continue to improve performance on these and related tasks.
    - Though we don't know how to scale up other universal approximators.
  - There is some overfitting to popular datasets like ImageNet.
    - Recent work showed accuracy drop of 11-14%by using a different ImageNet test set.

- CNNs are now making their way into products.
  - Face/person recognition in various cameras.
  - Car recognition in vehicles.
  - Amazon Go: https://www.youtube.com/watch?v=NrmMk1Myrxc
    - Trolling by French company Monoprix here.

Mission Accomplished?

- We're still missing a lot of theory and understanding deep learning.

From: Boris
To: Ali

On Friday, someone on another team changed the default rounding mode of some Tensorflow internals (from truncation to "round to even").*

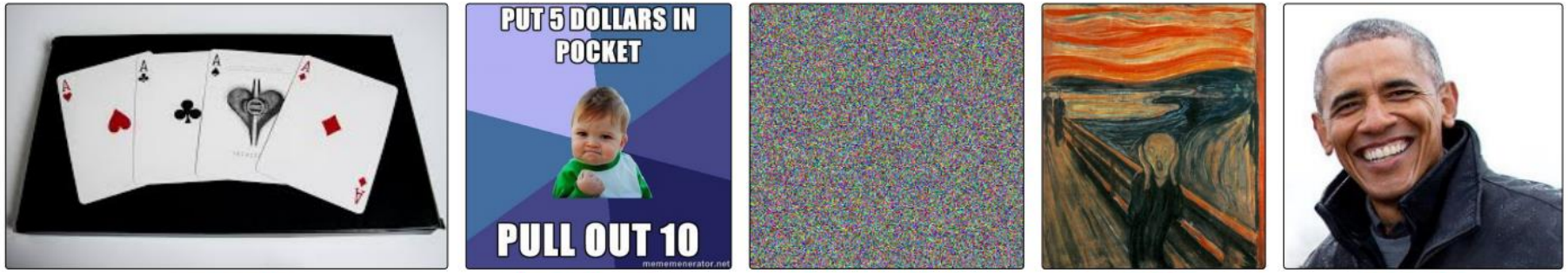
*Our training broke. Our error rate went from <25% error to ~99.97% error (on a standard 0-1 binary loss).

- "Good CS expert says: Most firms that think they want advanced AI/ML really just need linear regression on cleaned-up data."

http://www.argmin.net/2017/12/05/kitchen-sinks/

Mission Accomplished?

- Despite high-level of abstraction, deep CNNs are easily fooled:
  - What happens when you give a weird input to a CNN?

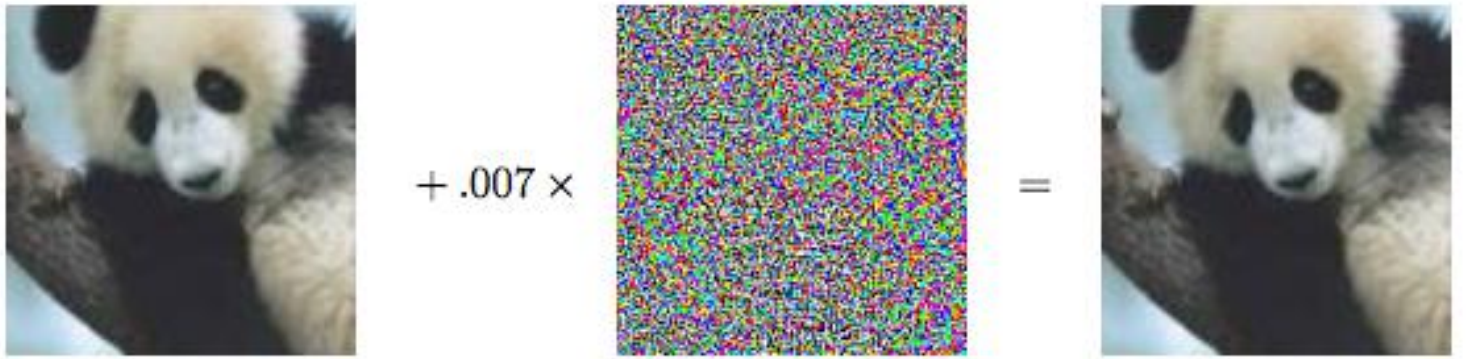


DenseNet 161 (2017)	Envelope 31%	Balance Beam 52%	Chainlink Fence 31%	Chest 37%	Tench 36%
SqueezeNet (2016)	Binder 43%	Balance Beam 18%	Poncho 32%	Jean 30%	Suit 21%
ResNet 152 (2015)	Envelope 40%	Pacifier 33%	Chain Mail 29%	Dust Cover 52%	Sweatshirt 25%
VGG 19 (2014)	Binder 51%	Dust Cover 44%	Window Screen 5%	Chest 11%	Sweatshirt 46%
AlexNet (2012)	T-shirt 16%	Dust Cover 22%	Cardigan 12%	Theater Curtain 3%	Coho 37%

Figure 1: The arbitrary predictions of several popular networks [2, 3, 4, 5, 6] that are trained on ImageNet [1] on unseen data. The red predictions are entirely wrong, the green predictions are justifiable, the orange predictions are less justifiable. The middle image is noise sampled from $\mathcal{N}(\mu = 0.5, \sigma = 0.25)$ without any modifications. This unpredictable behaviour is not limited to demonstrated architectures. We show that merely thresholding the output probability is not a reliable method to detect these problematic instances.

Mission Accomplished?

- Despite high-level of abstraction, deep CNNs are easily fooled:
  - What happens when you give a weird input to a CNN?
- Recent work: imperceptible noise that changes the predicted label.
  - "Adversarial examples" (can change to any other label).



https://arxiv.org/pdf/1412.6572.pdf, https://blog.openai.com/adversarial-example-research/

Mission Accomplished?

- Can someone repaint a stop sign and fool self-driving cars?

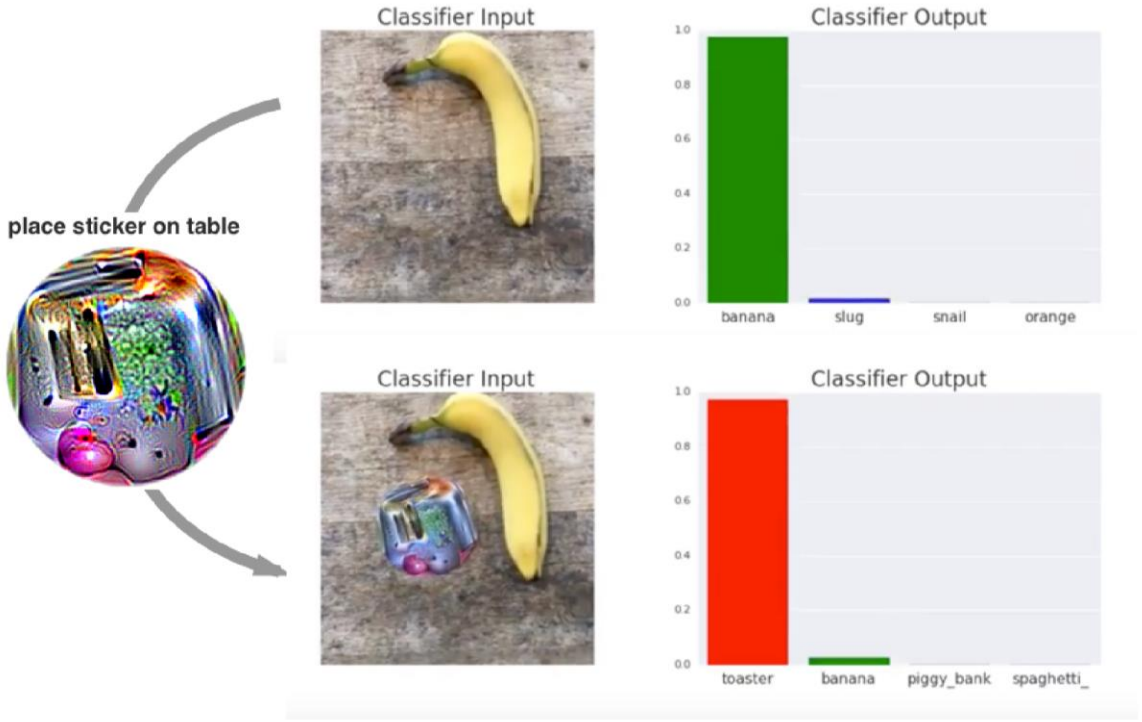


Figure 1: A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we physically place a sticker targeted to the class "toaster" on the table (bottom photograph), the photograph is classified as a toaster with 99% confidence (bottom plot). See the following video for a full demonstration: `https://youtu.be/i1sp4X57TL4`

https://arxiv.org/pdf/1712.09665.pdf

Mission Accomplished?

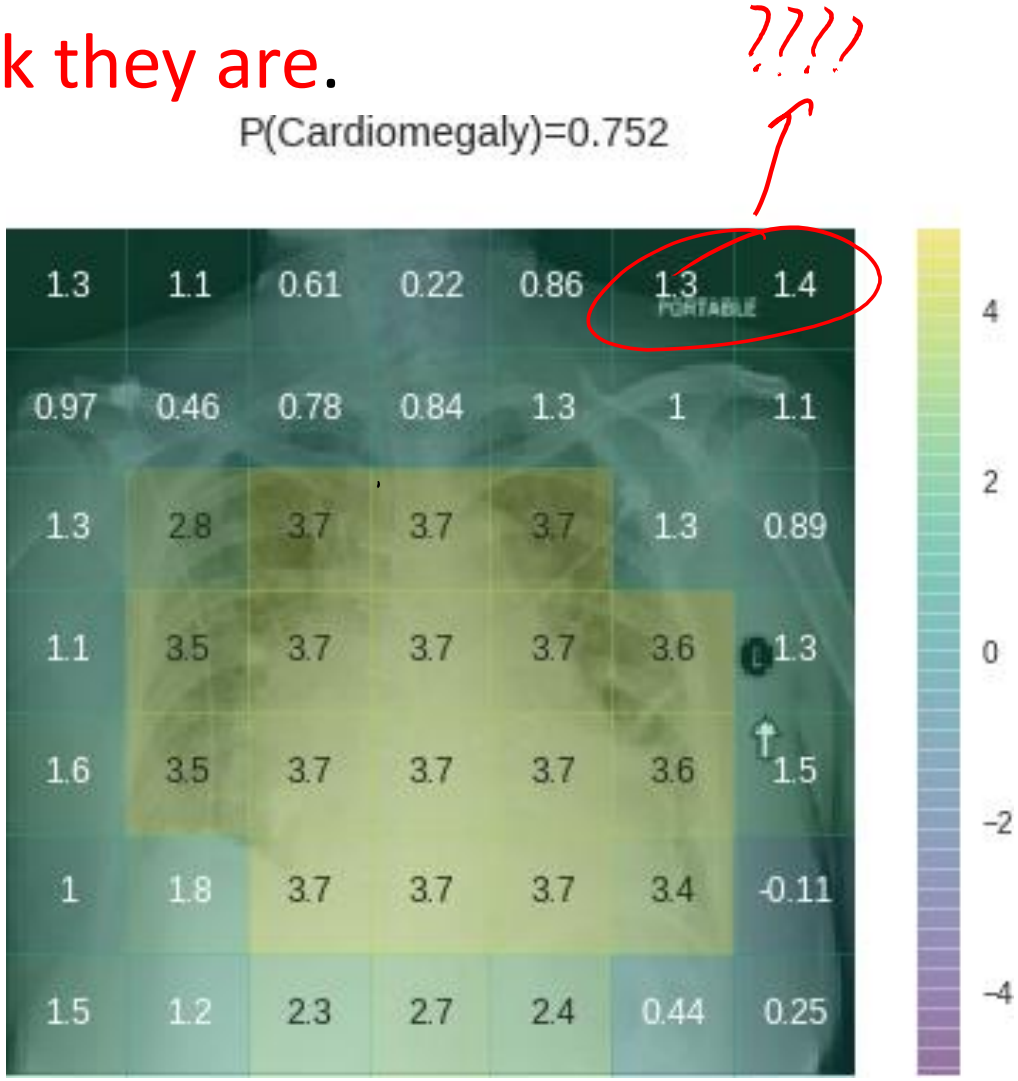
- Are the networks understanding the fundamental concepts?
  - Is being “surrounded by green” part of the definition of cow?
  - Do we need to have examples of cows in different environments?
    - Kids don’t need this.

- Image colourization:

https://www.onegreenplanet.org/news/cows-enjoy-the-beach/
https://mathwithbaddrawings.com/2017/10/18/5-ways-to-troll-your-neural-network/

Mission Accomplished?

- CNNs may not be learning what you think they are.
  - CNN for diagnosing enlarged heart:
    - Higher values mean more likely to be enlarged:
  - CNN says “portable” protocal is predictive:
    - But they are probabaly getting a “portable” scan because they’re too sick to go the hospital.
  - CNN was biased by the scanning protocal.
    - Learns the scans that more-sick patients get.
    - This is not what we want in a medical test.



https://medium.com/@jrzech/what-are-radiological-deep-learning-models-actually-learning-f97a546c5b98

Meaningless comparisons lead to false optimism in medical machine learning

Orianna DeMasi, Konrad Kording, Benjamin Recht

(Submitted on 19 Jul 2017)

A new trend in medicine is the use of algorithms to analyze big datasets, e.g. using everything your phone measures about you for diagnostics or monitoring. However, these algorithms are commonly compared against weak baselines, which may contribute to excessive optimism. To assess how well an algorithm works, scientists typically ask how well its output correlates with medically assigned scores. Here we perform a meta-analysis to quantify how the literature evaluates their algorithms for monitoring mental wellbeing. We find that the bulk of the literature (∼77%) uses meaningless comparisons that ignore patient baseline state. For example, having an algorithm that uses phone data to diagnose mood disorders would be useful. However, it is possible to over 80% of the variance of some mood measures in the population by simply guessing that each patient has their own average mood - the patient-specific baseline. Thus, an algorithm that just predicts that our mood is like it usually is can explain the majority of variance, but is, obviously, entirely useless. Comparing to the wrong (population) baseline has a massive effect on the perceived quality of algorithms and produces baseless optimism in the field. To solve this problem we propose "user lift" that reduces these systematic errors in the evaluation of personalized medical monitoring.

- Related: does the prediction change real-world outcomes?
  - Are you just annoying the highly-paid doctor or paying for nothing?

Racially-Biased Algorithms?

- Major issue: are we learning representations with harmful biases?
  - Biases could come from data (if data only has certain groups in certain situations).
  - Biases could come from labels (always using label of “ball” for certain sports).
  - Biases could come from learning method (model predicts “basketball” for black people more often than this appears in training data for basketball images).

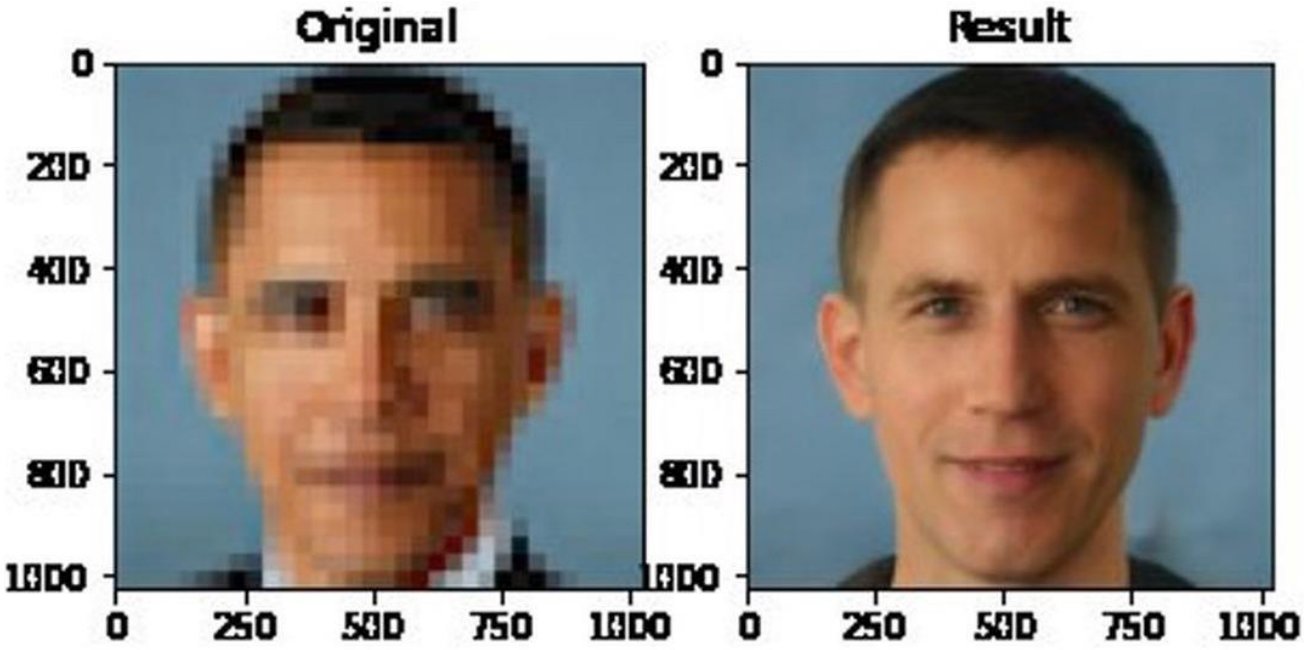
Fig. 8: Pairs of pictures (columns) sampled over the Internet along with their prediction by a ResNet-101.

  - This is a major problem/issue when deploying these systems.
    - For example, “repeat-offender prediction” that reinforces racial biases in arrest patterns.

https://arxiv.org/pdf/1711.11443.pdf

Racially-Biased Algorithms?

- Results on image super-resolution (upscaling) method:

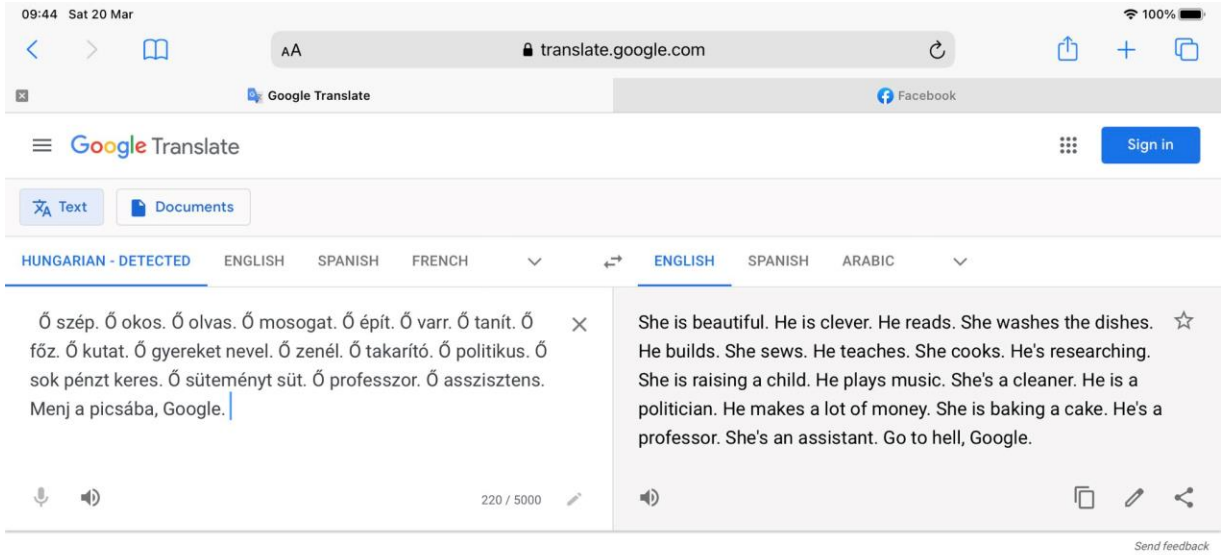


- Sometimes these issues can be reduced by careful data collection.
  - In this case, we could train on a more-diverse group.
  - But sometimes you cannot collect unbiased data.

https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

Sexist Algorithms?

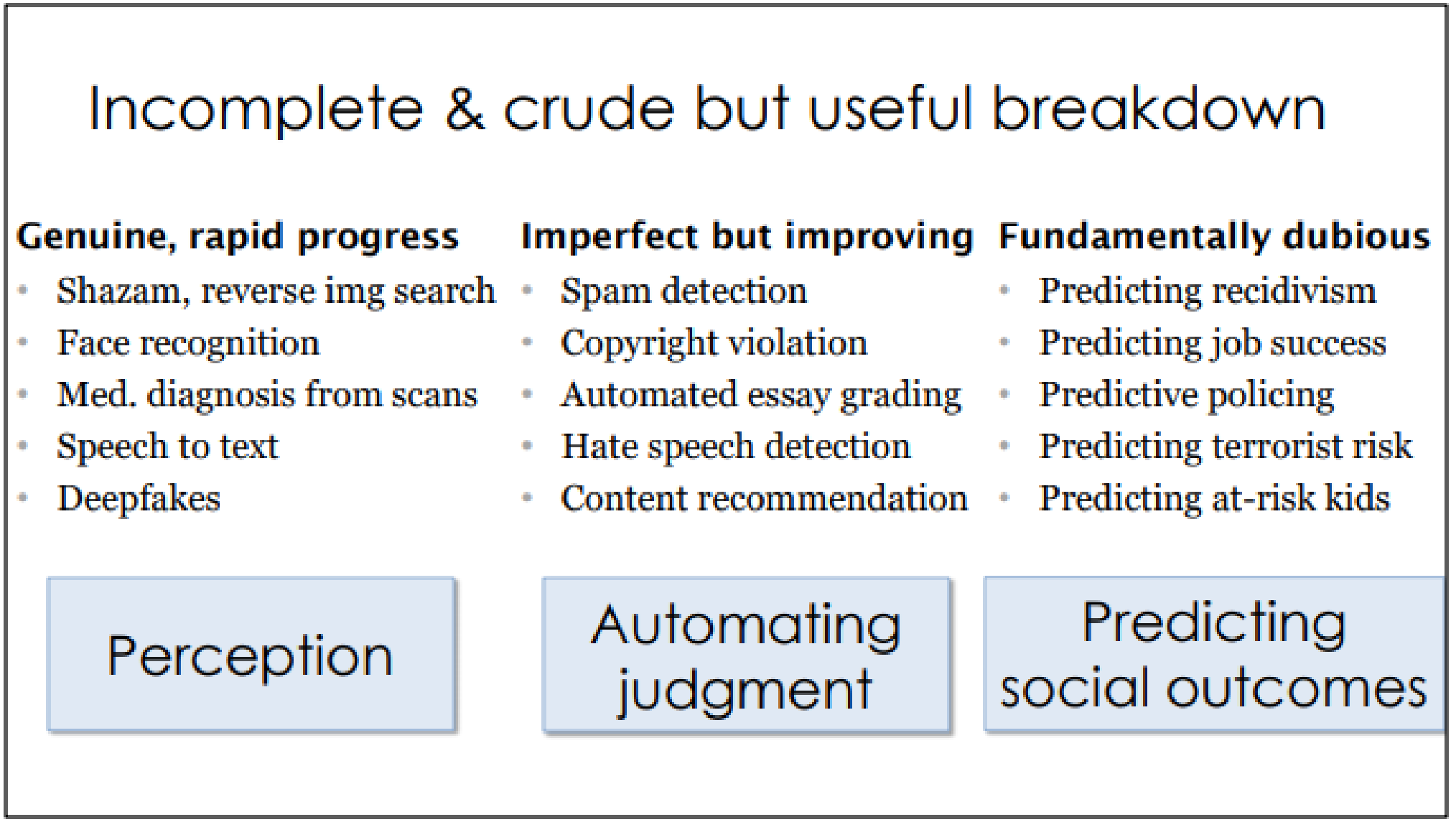
- Hungarian is gender neutral.
  - Google assigns a gender based on frequencies in training set:



- Amazon's hiring algorithm penalized candidates with "woman/women" in application.
  - Another correlation/causation issue: "most engineers at Amazon are men, engineers should be men?"
- Maybe we will eventually fix issues like this.
  - Until we do, maybe we should not use machine learning in some applications.
    - Or at least warn people about potential biases.

https://twitter.com/doravargha/status/1373211762108076034

- From "How to Recognize AI Snake Oil".
  - https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf



Summary

- CNNs seem to be learning sensible things.
  - Earlier layers seem to represent low-level features.
  - Later layers seem to represent complex object-level features.
- But our ML models are easily fooled:
  - Unpredictable performance on “out of distribution” images.
  - Adversarial examples lead to incorrect predictions.
  - Do not understand context and obvious confounding factors.
- There are some problems where we should not use ML.
  - ML models can learn/propagate/enhance harmful biases (sometimes fixable, sometimes not).
  - On some applications, methods do not work better than obvious baselines.

- Next time: supervised learning where we do not know the output size.