CPSC 340: Machine Learning and Data Mining

More CNNs

Fall 2019

AlexNet Convolutional Neural Network

- ImageNet 2012 won by AlexNet:
  - 15.4% error vs. 26.2% for closest competitor.
  - 5 convolutional layers.
  - 3 fully-connected layers.
  - SG with momentum.
  - ReLU non-linear functions.
  - Data translation/reflection/ cropping.
  - L2-regularization + Dropout.
  - 5-6 days on two GPUs.
  - Same networks won in 2013: tweaks like smaller stride and smaller filters.

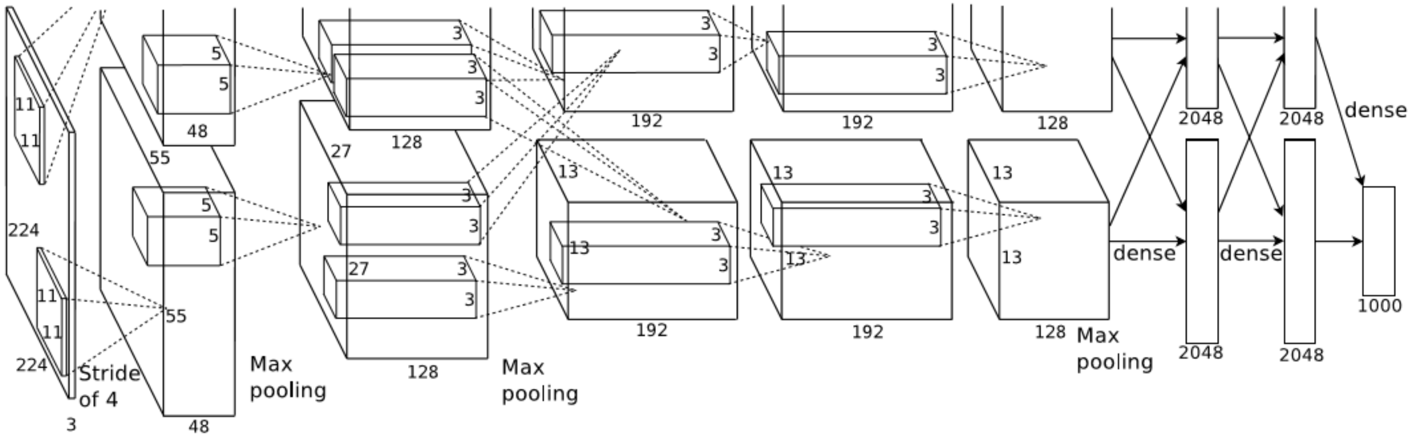
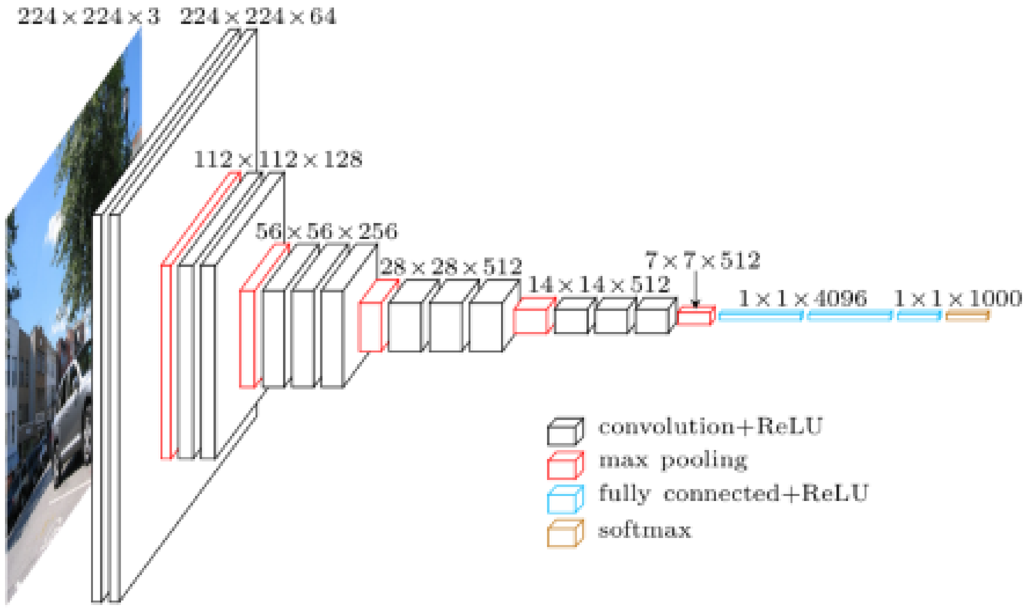


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input is 150,528-dimensional, and the number of neurons in the network's remaining layers is given by 253,440–186,624–64,896–64,896–43,264–4096–4096–1000.

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

ImageNet Insights

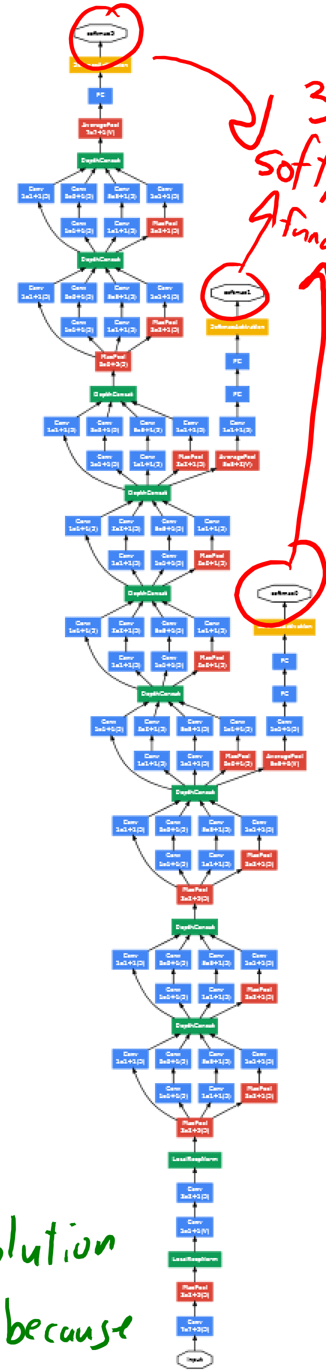
- Filters and stride got smaller over time.
  - Popular VGG approach uses 3x3 convolution layers with stride of 1.
    - 3x3 followed by 3x3 simulates a 5x5, and another 3x3 simulates a 7x7, and so on.
    - Speeds things up and reduces number of parameters.
    - Increases number of non-linear ReLU operations.



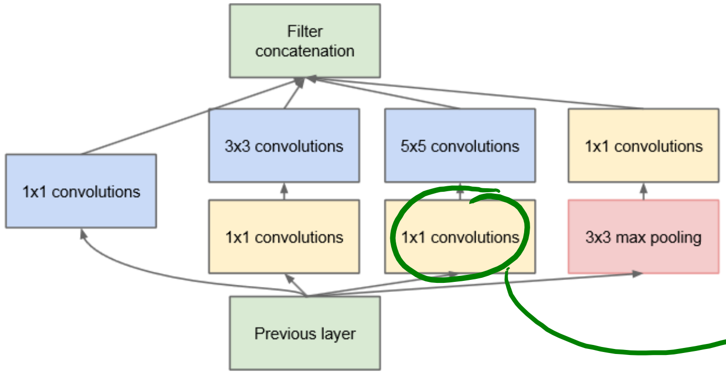
https://www.cs.toronto.edu/~frossard/post/vgg16/

ImageNet Insights

- Filters and stride got smaller over time.
  - Popular VGG approach uses 3x3 convolution layers with stride of 1.
  - GoogLeNet considered multiple filter sizes, but not as popular.
- Eventual switch to "fully-convolutional" networks.
  - No fully connected layers.



3 softmax functions



(b) Inception module with dimensionality reduction

"1x1" convolution makes sense because these are first 2 dimensions of 3D conv

https://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf
http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

ImageNet Insights

- Filters and stride got smaller over time.
  - Popular VGG approach uses 3x3 convolution layers with stride of 1.
  - GoogLeNet considered multiple filter sizes, but not as popular.
- Eventual switch to "fully-convolutional" networks.
  - No fully connected layers.
- ResNets allow easier training of deep networks.
  - Won all 5 tasks in 2015, training 152 layers for 2-3 weeks on 8 GPUs.
- Ensembles help.
  - Combine predictions of previous networks.

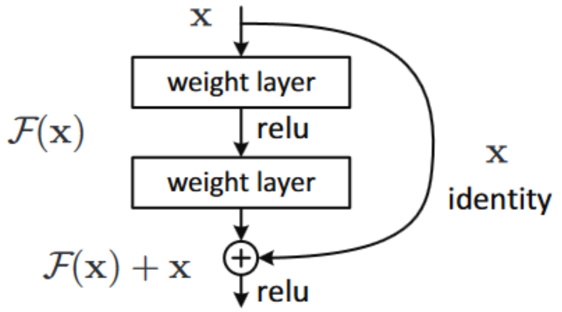
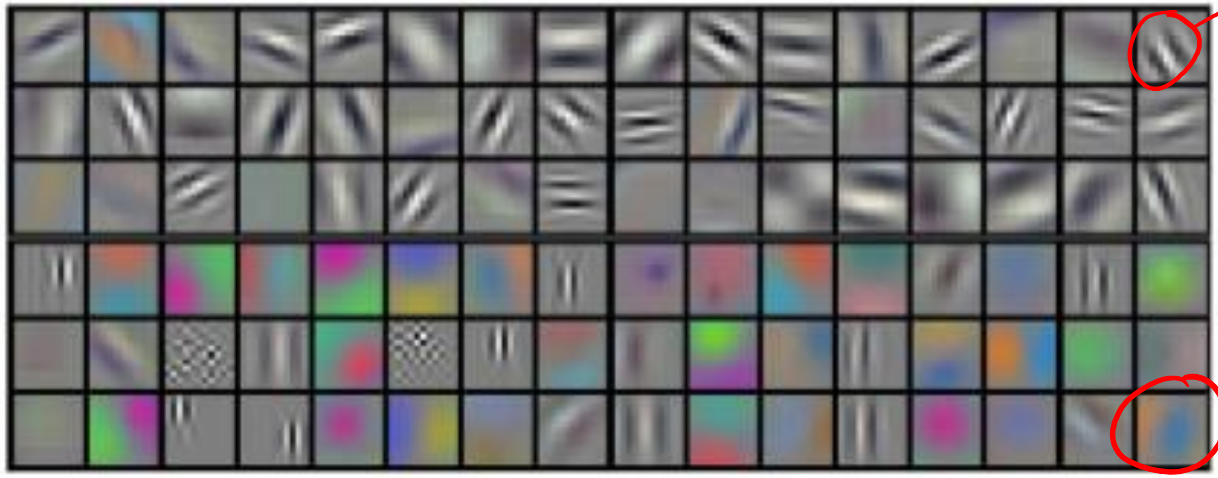


Figure 2. Residual learning: a building block.

Are CNNs learning something sensible?

- Filters learned by first layer of original AlexNet:



"Gabor" filters:
- Gaussian times sine or cosine.

"Opponent" colour coding.

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

- Note that non-orthogonal PCA gives similar results (but only 1 layer).

Are CNNs learning something sensible?

- It's harder to visualize what is learned in other layers.
  - Deconvolution networks try to reconstruct what "activates" filters.

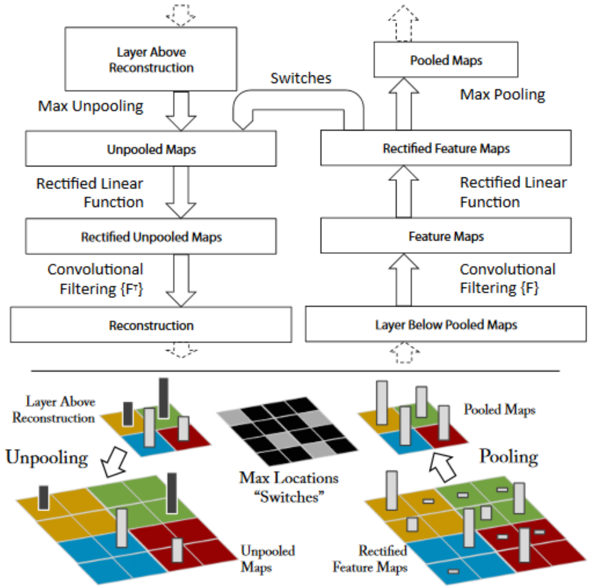
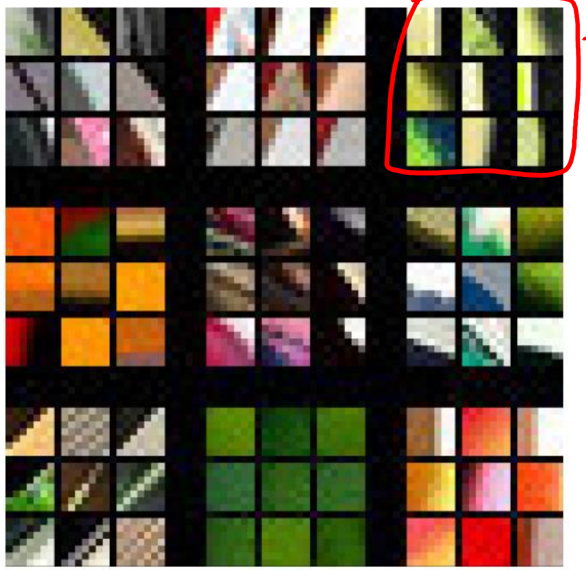


Figure 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local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

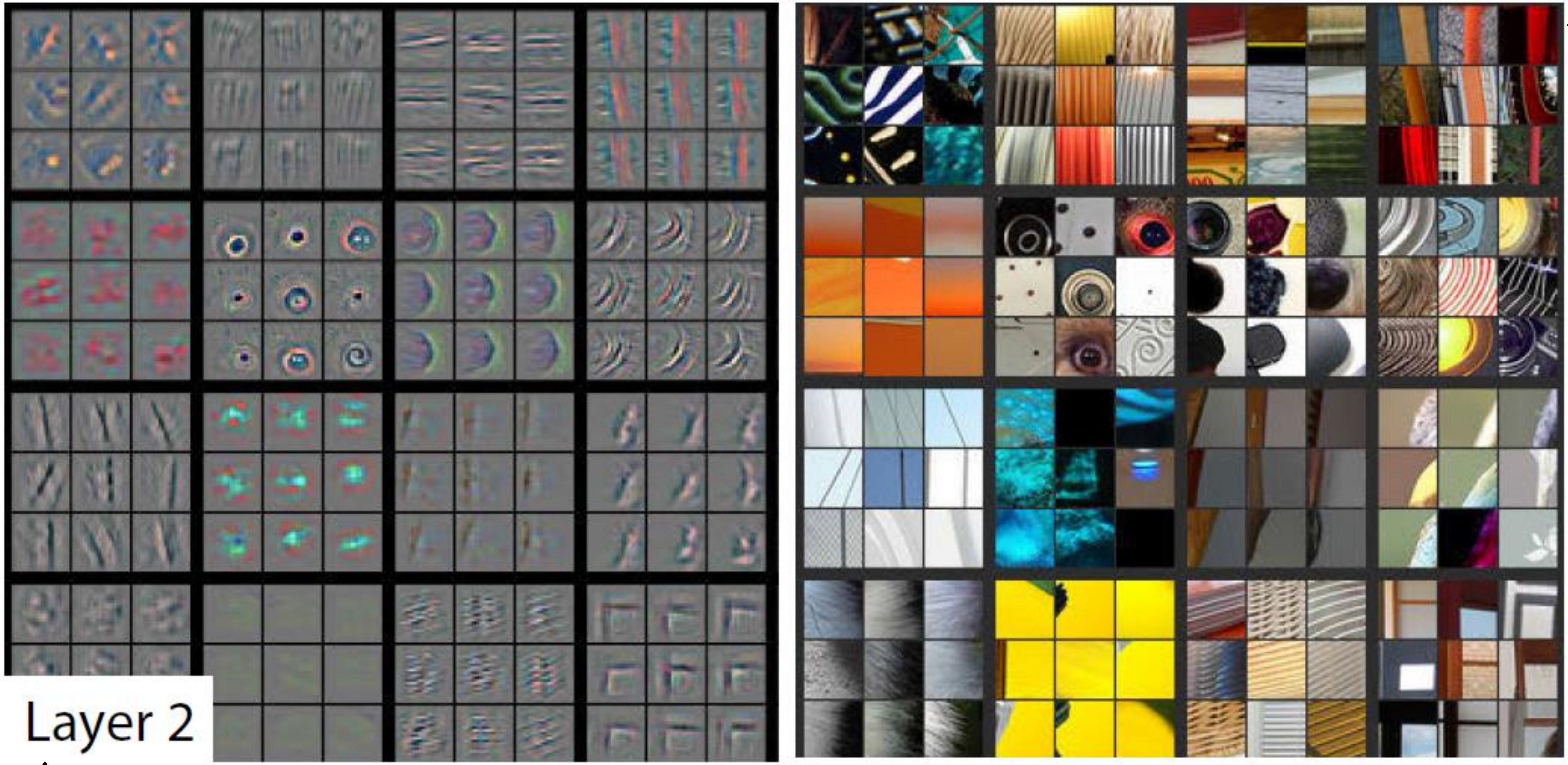
Layer 1

Synthetic image that makes a filter respond most strongly.



Real images where filter responds most strongly.

Are CNNs learning something sensible?

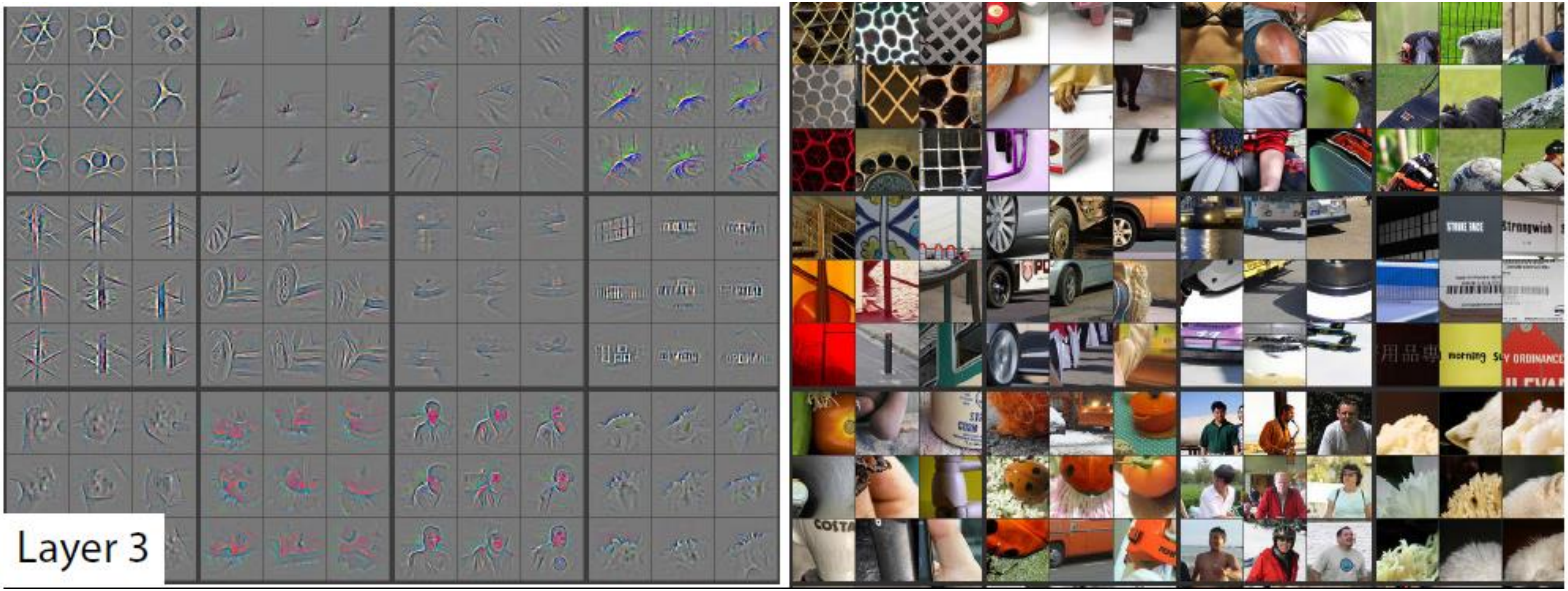


Patch from data giving largest response

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

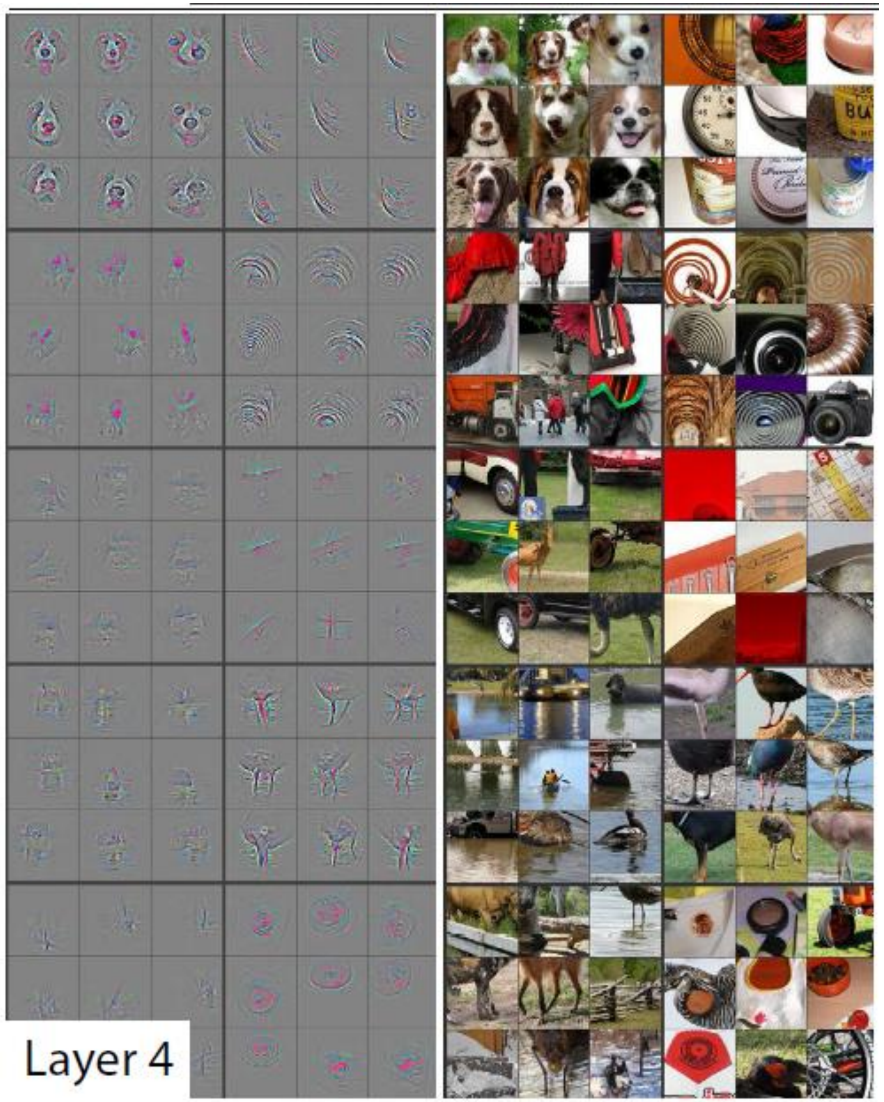
Deconvolution network giving patch that leads to largest response

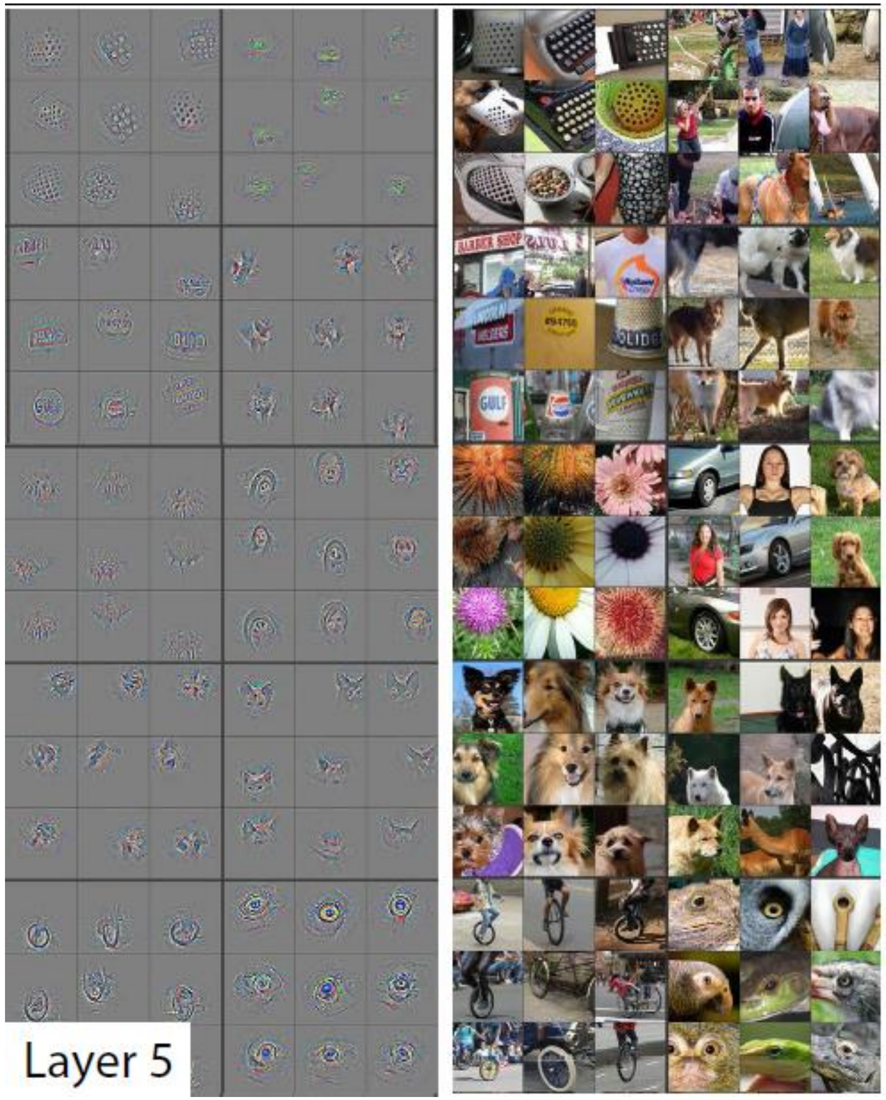
Are CNNs learning something sensible?



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

Are CNNs learning something sensible?

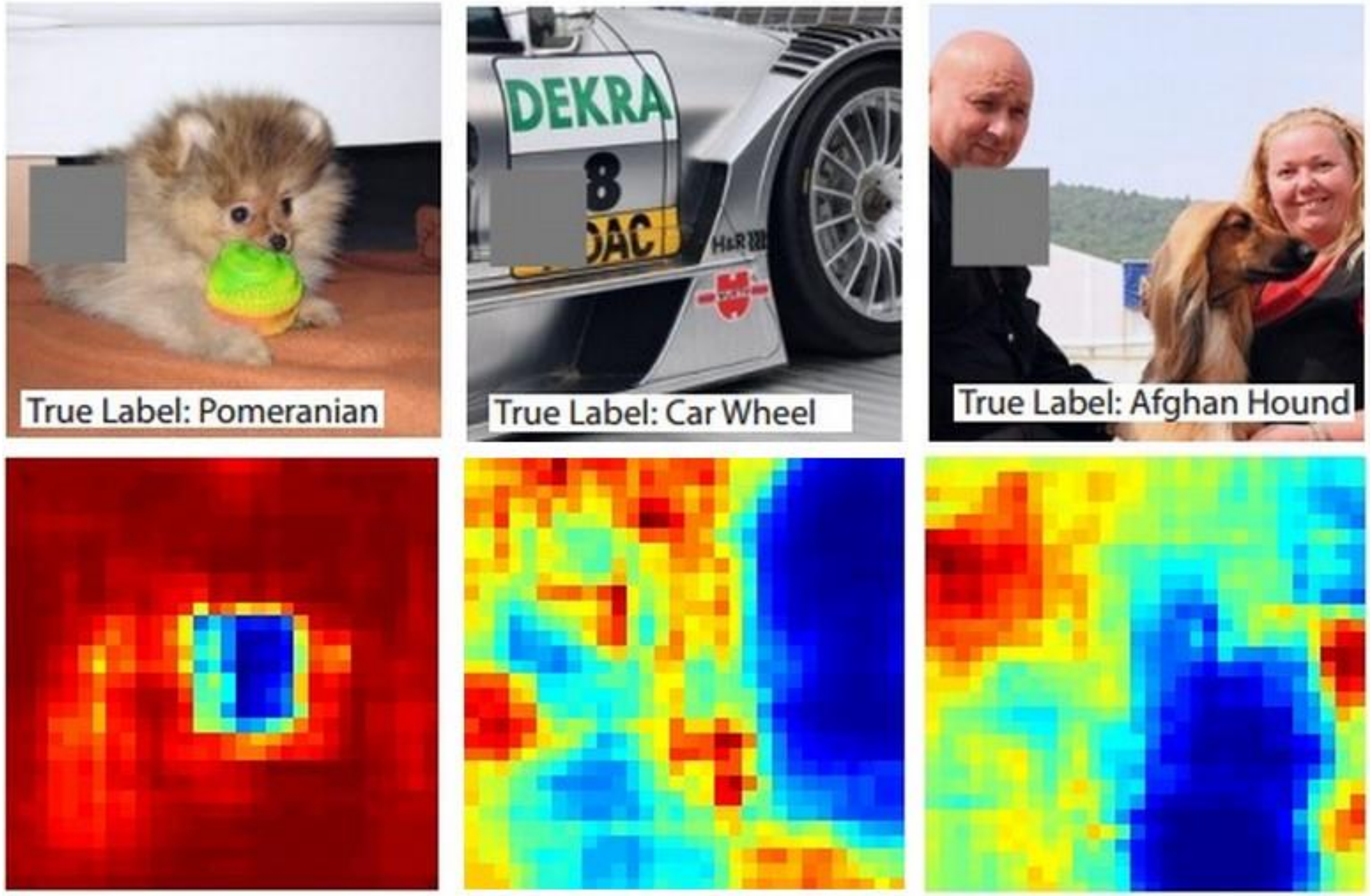




https://arxiv.org/pdf/1311.2901v3.pdf

Are CNNs learning something sensible?

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



http://cs231n.github.io/understanding-cnn/

Mission Accomplished?

- For speech recognition and object detection:
  - No other methods have ever given the current level of performance.
  - Deep models continue to improve performance on these and related tasks.
  - We don't know how to scale up other universal approximators.
  - There is likely some overfitting to popular datasets like ImageNet.
    - Recent work showed accuracy drop of 4-10% by using a different test set on CIFAR 10.

- CNNs are now making their way into products.
  - Face recognition.
  - Amazon Go: https://www.youtube.com/watch?v=NrmMk1Myrxc
    - Trolling by French company Monoprix here.
  - Self-driving cars.

Mission Accomplished?

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

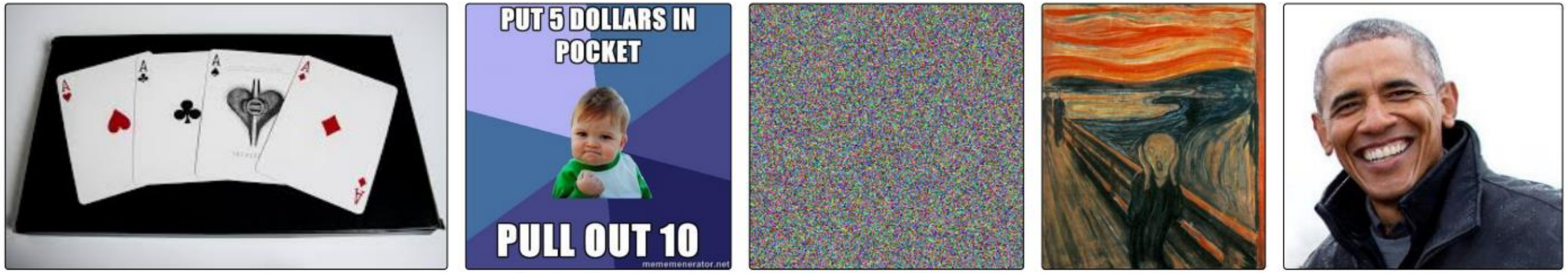
```
From: Boris
To: Ali

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 thinks 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:
  - Hot research topic at the moment.

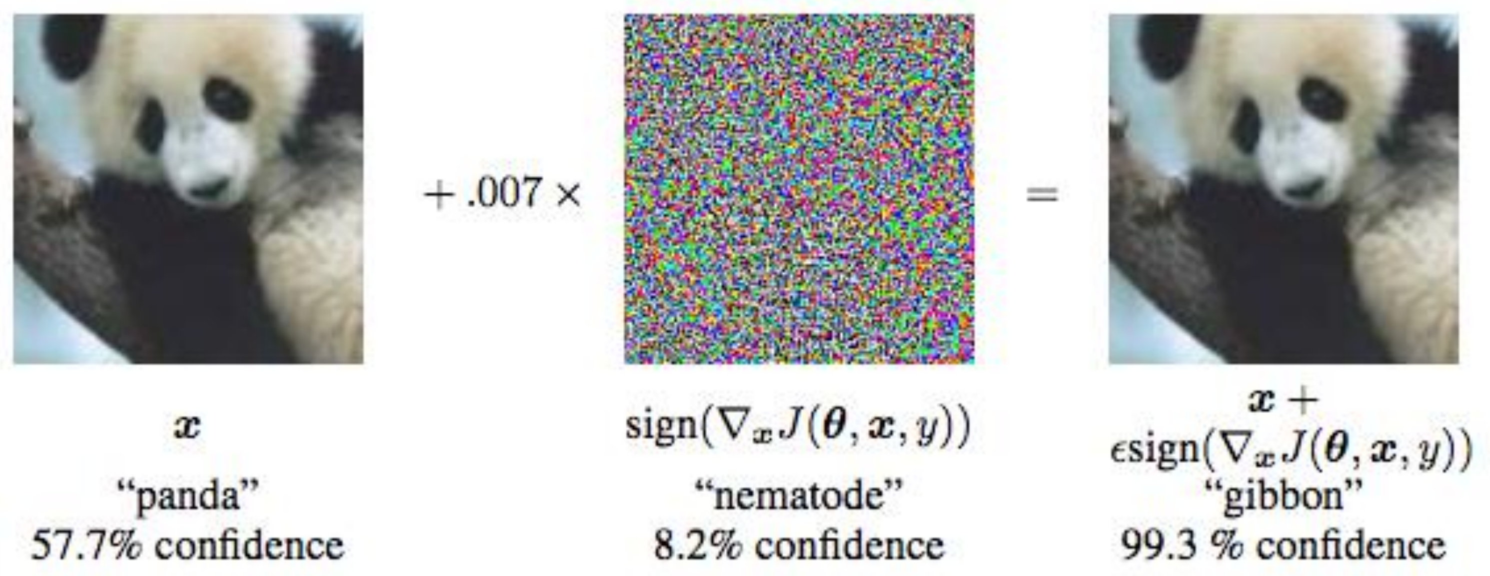


DenseNet 161 (2017)	Envelope 31%	Balance Beam 52%	Chainlink Fence 31%	Chest 37%	Tench 36%
SqueezeNet (2016)	Binder 43%	Balance Beam 18%	Poncho 32%	Jean 30%	Suit 21%
ResNet 152 (2015)	Envelope 40%	Pacifier 33%	Chain Mail 29%	Dust Cover 52%	Sweatshirt 25%
VGG 19 (2014)	Binder 51%	Dust Cover 44%	Window 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:
  - Hot research topic at the moment.
- 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?

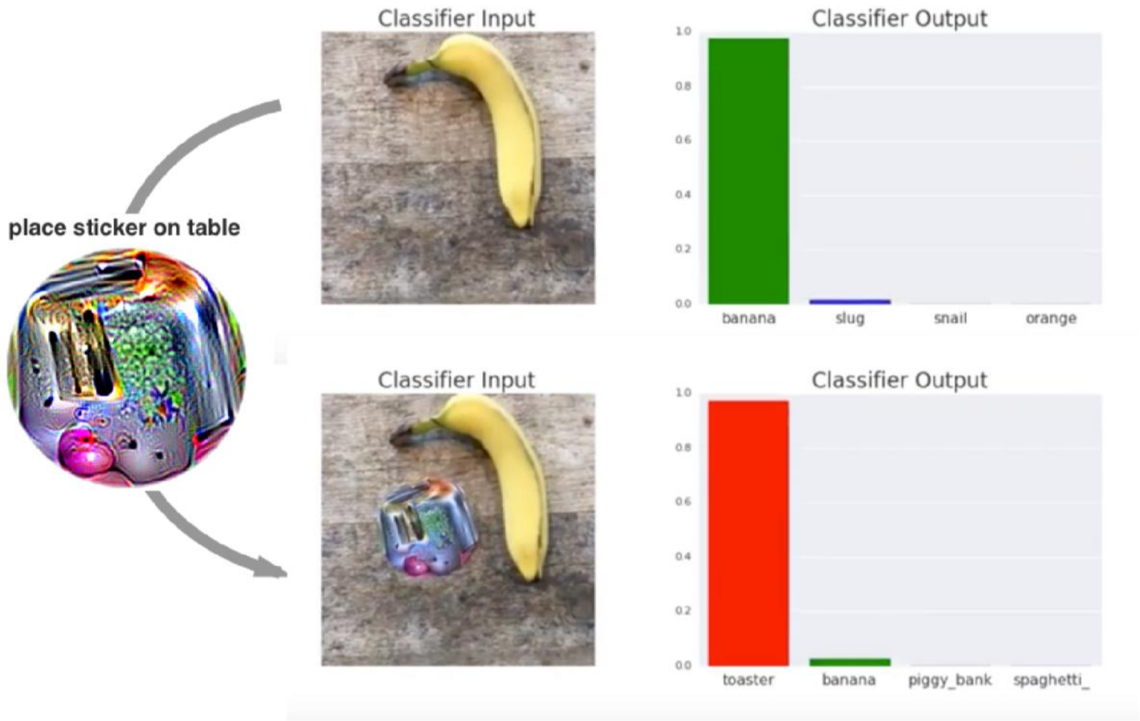


Figure 1: A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we 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

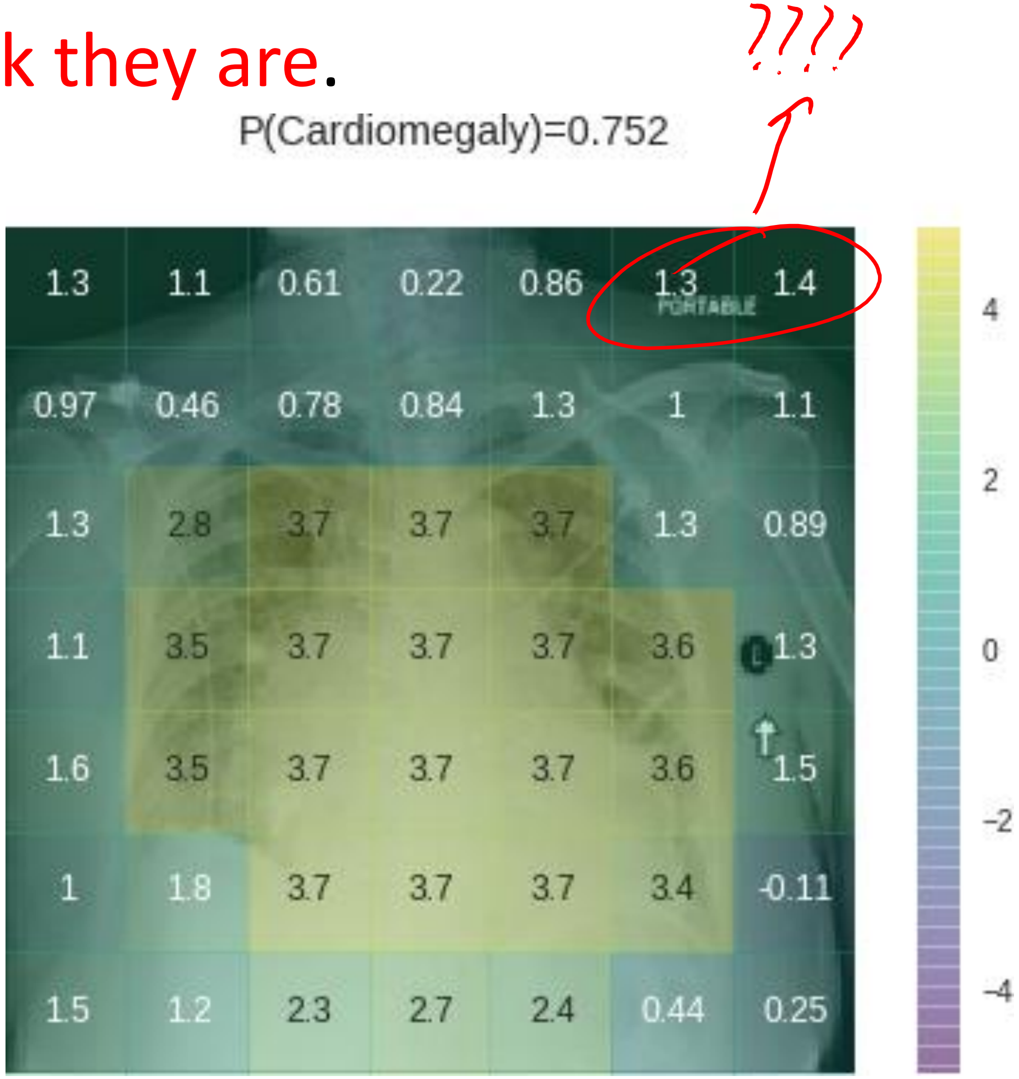
Mission Accomplished?

- 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.

https://www.onegreenplanet.org/news/cows-enjoy-the-beach/

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

(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 they appear 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.
    - E.g., “repeat-offender prediction” that reinforces racial biases in arrest patterns.

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

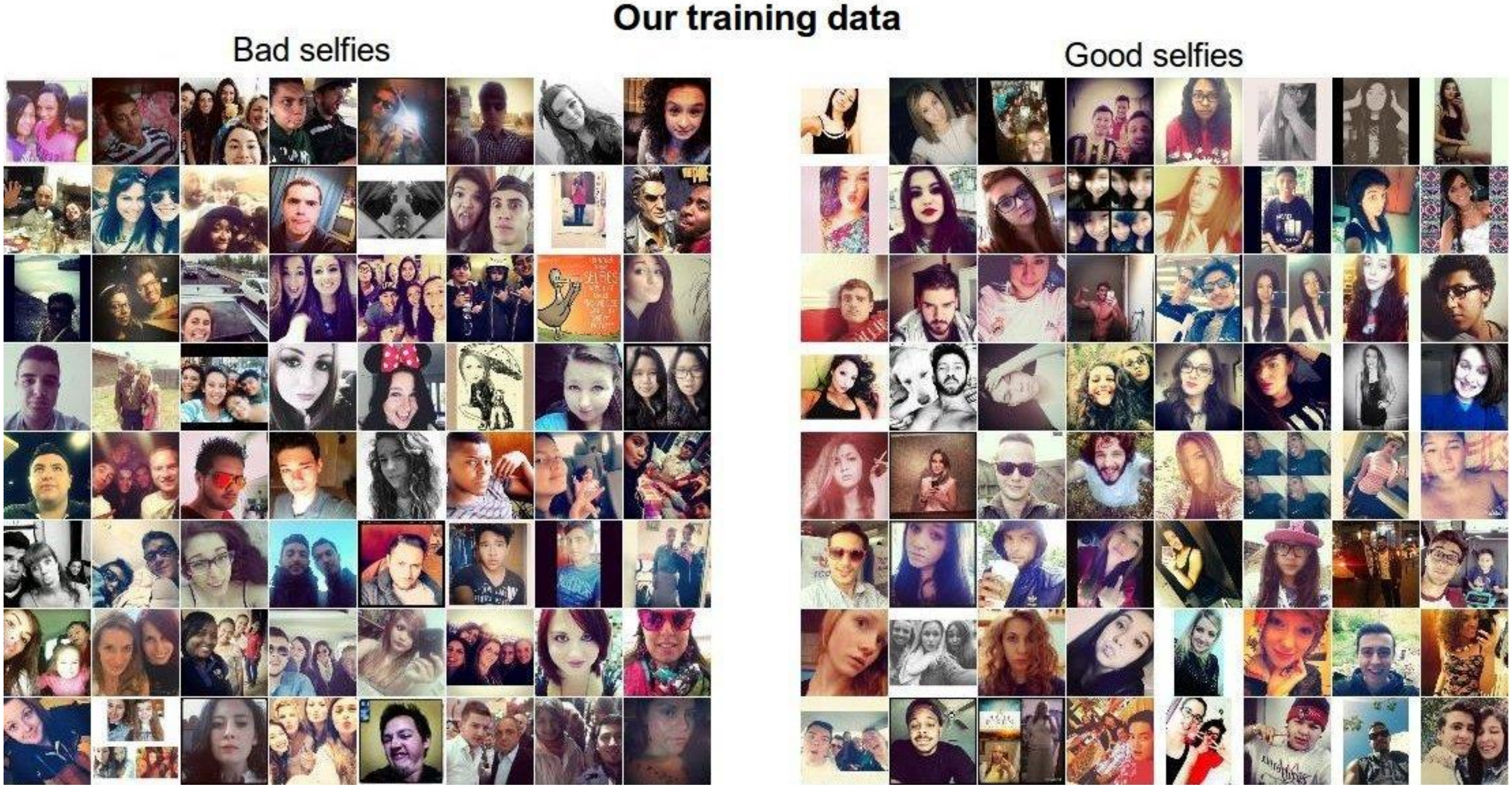
Energy Costs

- Current methods require:
  - A lot of data.
  - A lot of time to train.
  - Many training runs to do hyper-parameter optimization.

- Recent paper regarding recent deep language models:
  - Entire training procedure emits 5 times more CO_2 than lifetime emission of a car, including making the car.

(pause)

CNNs for Rating Selfies

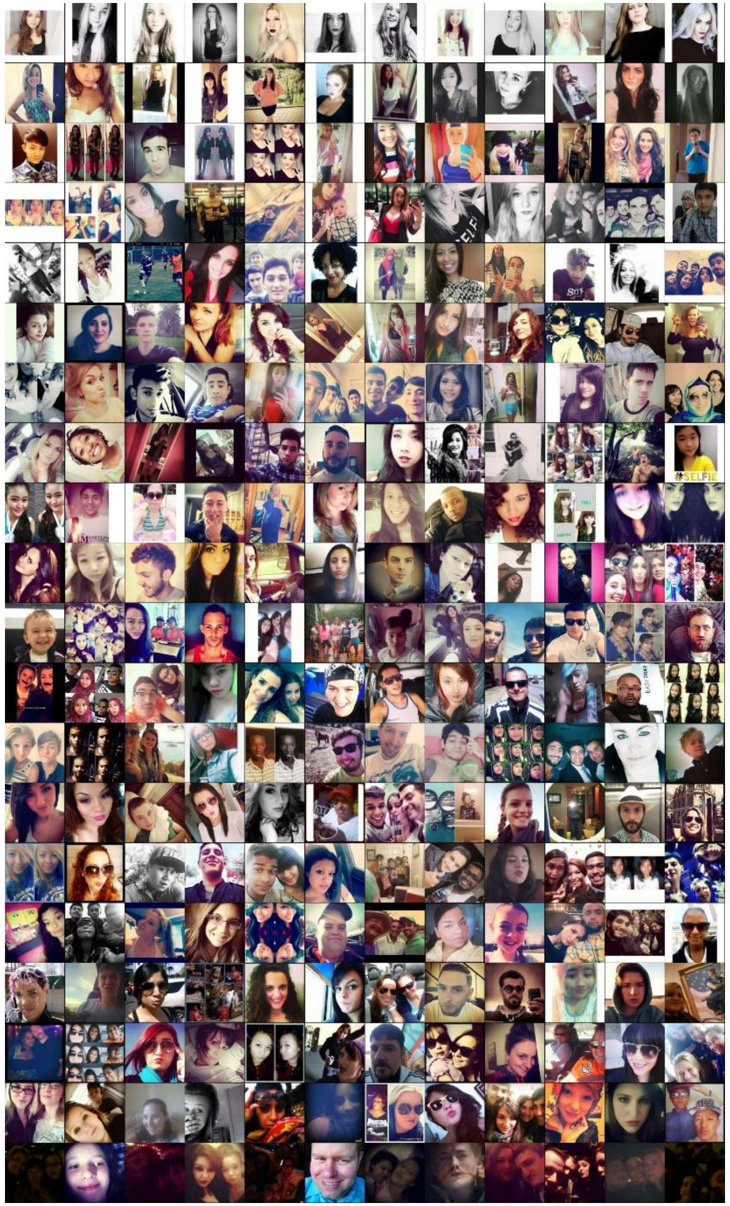


https://karpathy.github.io/2015/10/25/selfie/

CNNs for Rating Selfies

Do:

- Be female
- Have face be 1/3 of image
- Cut off forehead
- Show long hair
- Oversaturate face
- Use filter
- Add border



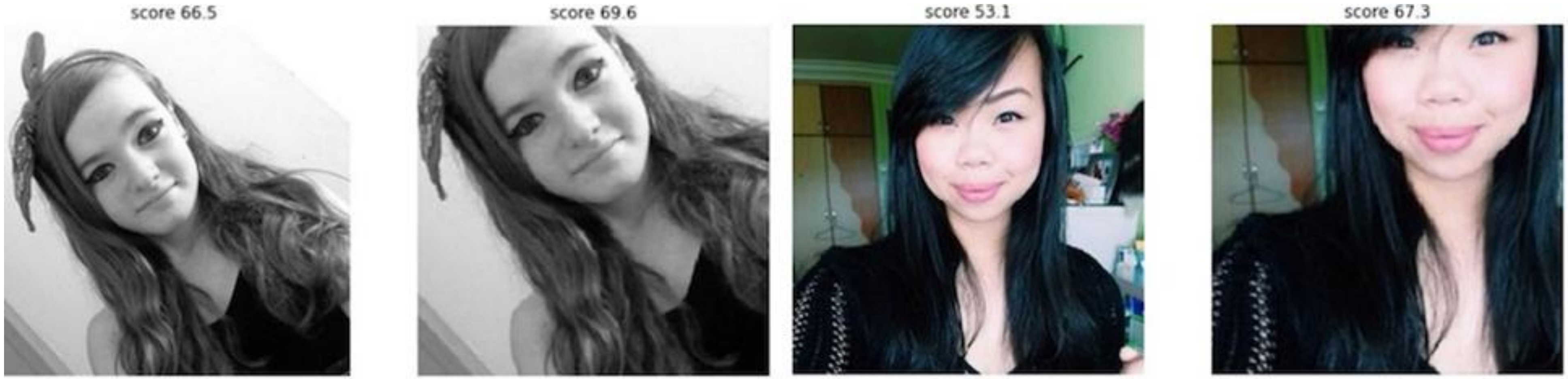
Don't:

- Use low lighting
- Make head too big
- Take group shots

https://karpathy.github.io/2015/10/25/selfie/

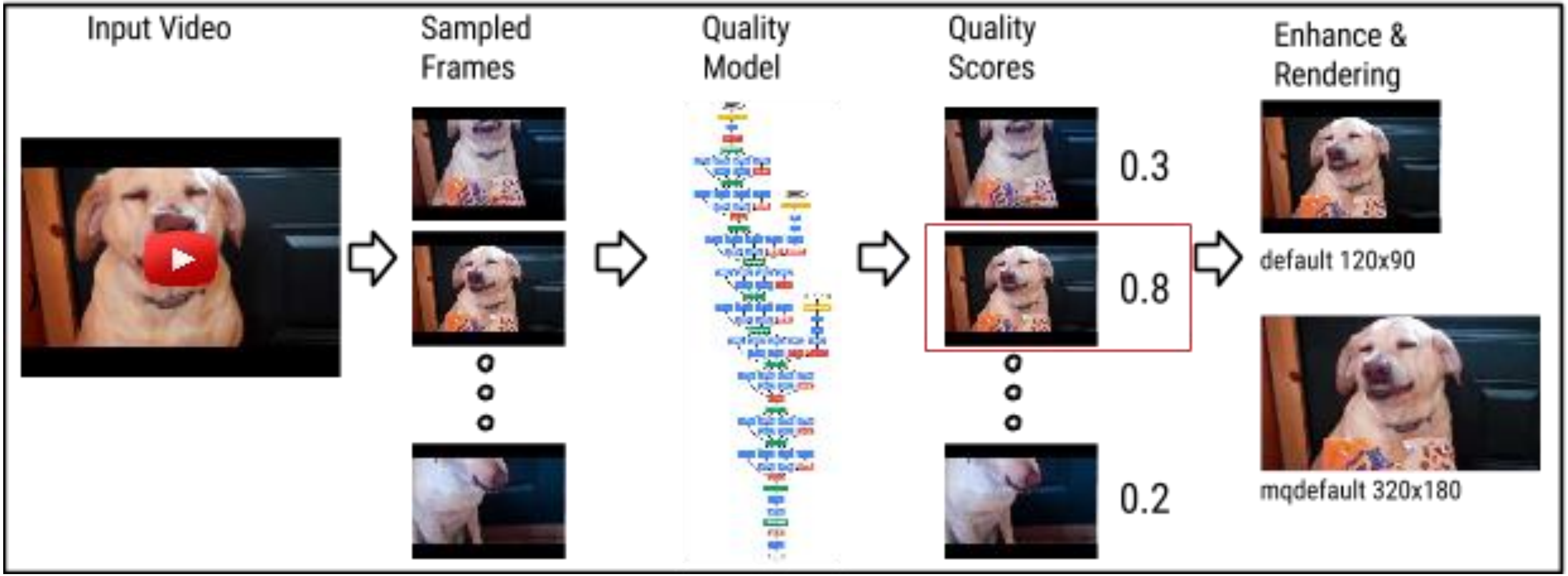
CNNs for Rating Selfies

Finding best image crop:



https://karpathy.github.io/2015/10/25/selfie/

CNNs for Choosing YouTube Thumbnails



https://youtube-eng.googleblog.com/2015/10/improving-youtube-video-thumbnails-with_8.html

Beyond Classification (CPSC 540)

- "Fully convolutional" neural networks allow "dense" prediction:

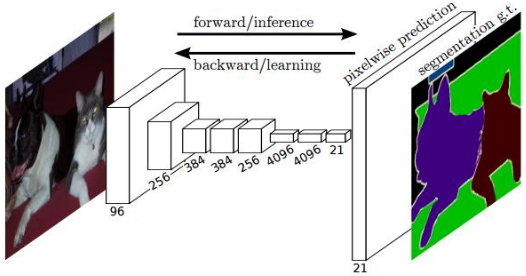


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

- Image segmentation:

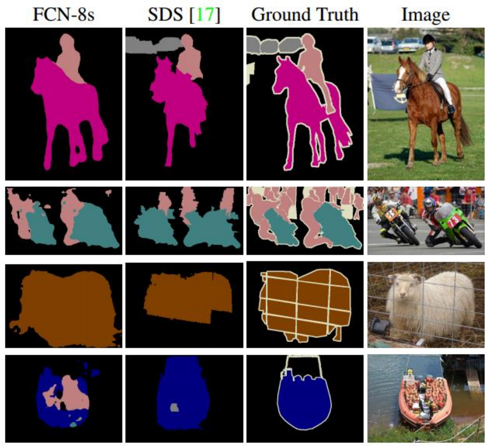
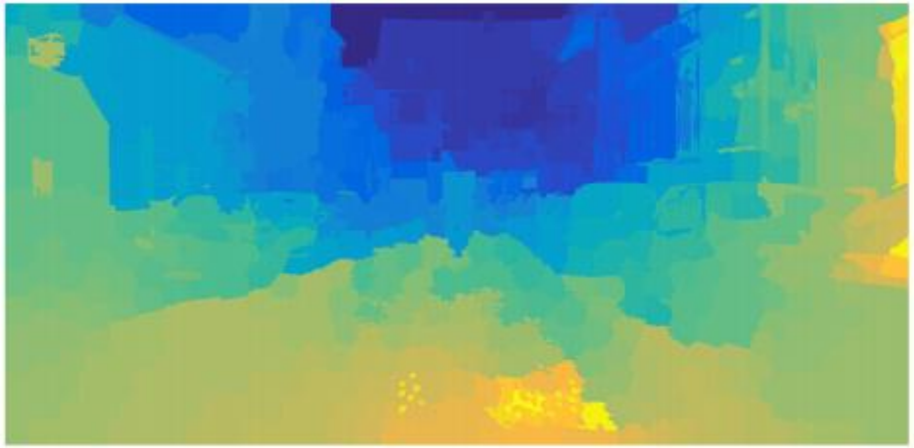


Figure 6. Fully convolutional segmentation nets produce state-of-the-art performance on PASCAL. The left column shows the output of our highest performing net, FCN-8s. The second shows the segmentations produced by the previous state-of-the-art system by Hariharan *et al*. [17]. Notice the fine structures recovered (first

https://people.eecs.berkeley.edu/~jonlong/long_shelhamer_fcn.pdf

Beyond Classification (CPSC 540)

- Depth Estimation:



- ["A Year in Computer Vision"]

Beyond Classification (CPSC 540)

- “AutoPortrait”: automatic photo re-touching.

Beyond Classification (CPSC 540)

- Image colorization:

Colorado National Park, 1941 | Textile Mill, June 1937 | Berry Field, June 1909 | Hamilton, 1936

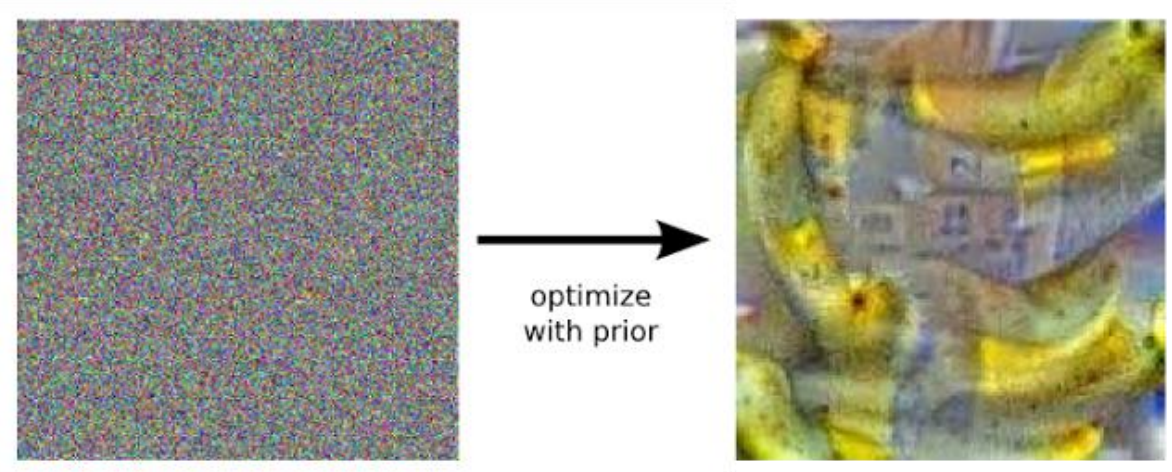
  - Image Gallery, Video

http://hi.cs.waseda.ac.jp/~iizuka/projects/colorization/en

Inceptionism

- A crazy idea:
  - Instead of weights, use backpropagation to take gradient with respect to x_i.
- Inceptionism with trained network:
  - Fix the label y_i (e.g., "banana").
  - Start with random noise image x_i.
  - Use gradient descent on image x_i.
  - Add a spatial regularizer on x_{ij}:
    - Encourages neighbouring x_{ij} to be similar.

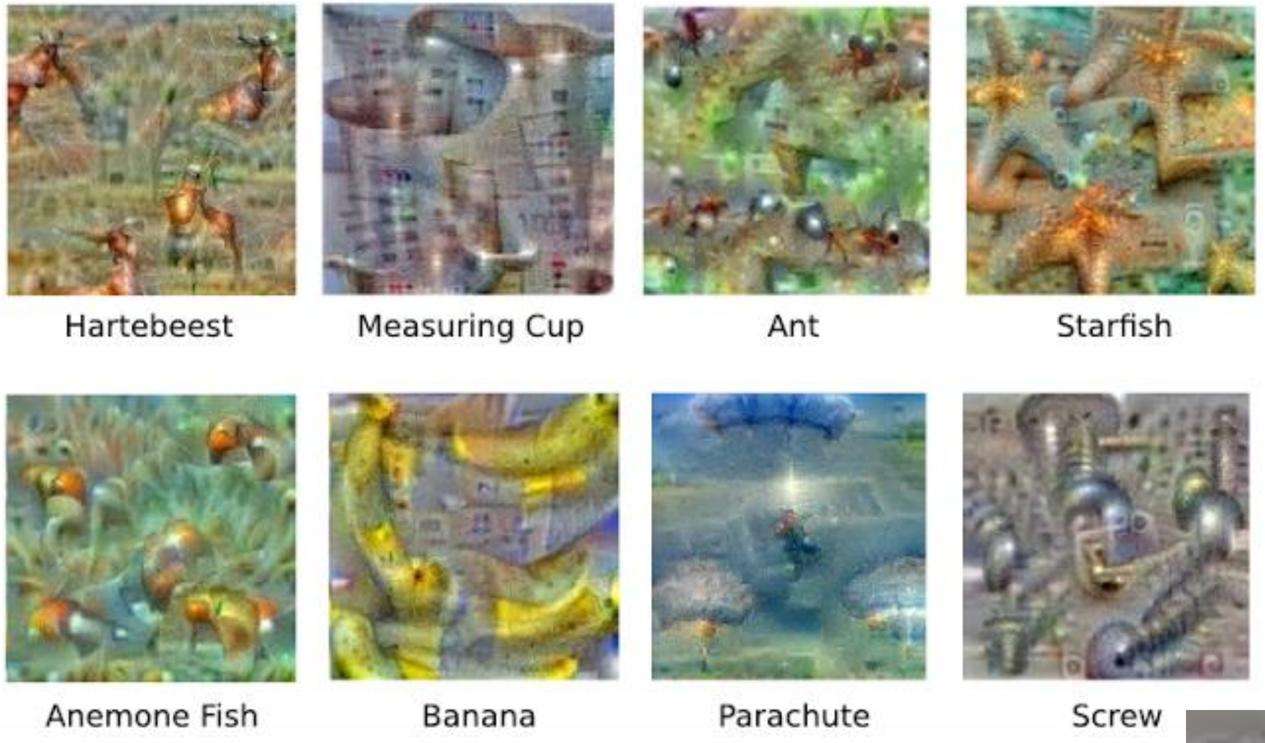
"Show what you think a banana looks like"



http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Inceptionism

- Inceptionism for different class labels:

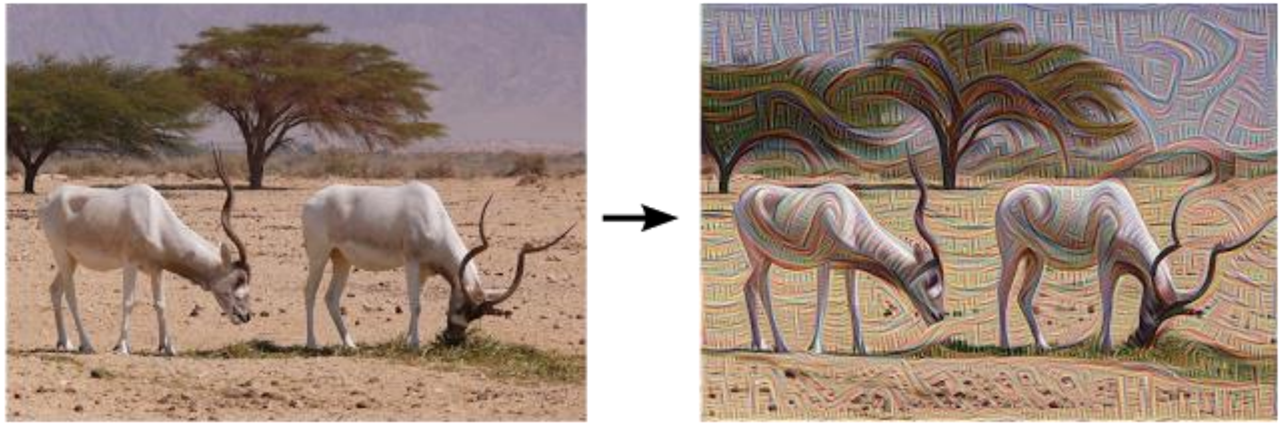


Dumbbell

http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

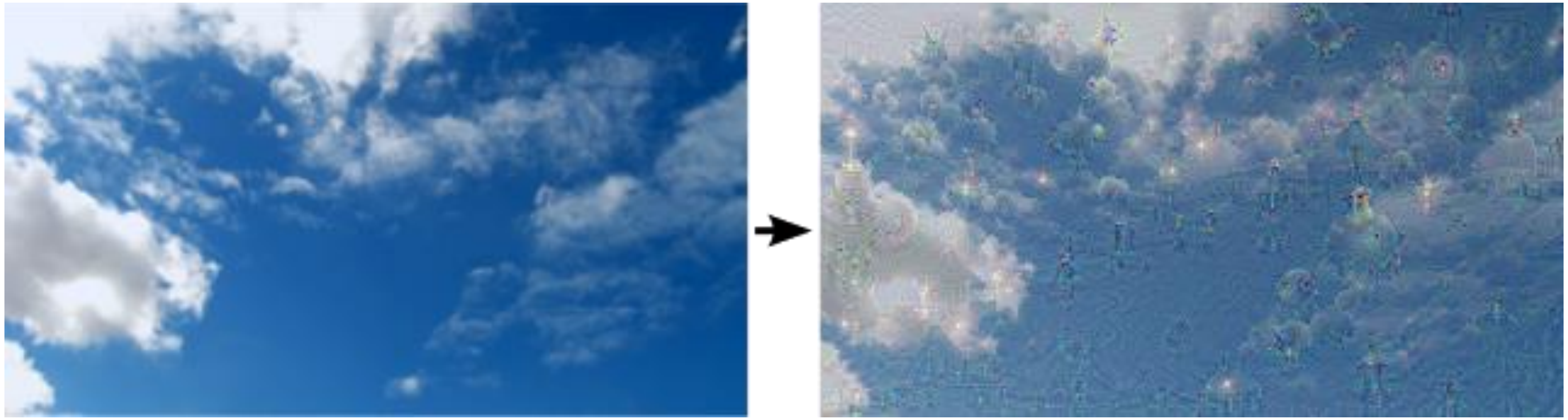
Inceptionism

- Inceptionism where we try to match $z_i^{(m)}$ values instead of y_i.
  - Shallow ‘m’:



Inceptionism

- Inceptionism where we try to match $z_i^{(m)}$ values instead of y_i.
  - Deepest ‘m’:



"Admiral Dog!"

"The Pig-Snail"

"The Camel-Bird"

"The Dog-Fish"

http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Inceptionism

- Inceptionism where we try to match $z_i^{(m)}$ values instead of y_i.
  - "Deep dream" starts from random noise:

  - Deep Dream video

http://googleresearch.blogspot.ca/2015/06/inceptionism-going-deeper-into-neural.html

Artistic Style Transfer

- Artistic style transfer:
  - Given a content image ‘C’ and a style image ‘S’.
  - Make a image that has content of ‘C’ and style of ‘S’.

Content:

Style:

https://commons.wikimedia.org/wiki/File:Tuebingen_Neckarfront.jpg
https://en.wikipedia.org/wiki/The_Starry_Night

Artistic Style Transfer

- Artistic style transfer:
  - Given a content image ‘C’ and a style image ‘S’.
  - Make a image that has content of ‘C’ and style of ‘S’.

- CNN-based approach applies gradient descent with 2 terms:
  - Loss function: match deep latent representation of content image ‘C’:
    - Difference between $z_i^{(m)}$ for deepest ‘m’ between x_i and ‘C’.
  - Regularizer: match all latent representation covariances of style image ‘S’.
    - Difference between covariance of $z_i^{(m)}$ for all ‘m’ between x_i and ‘S’.

Artistic Style Transfer

http://arxiv.org/pdf/1508.06576v2.pdf

Image Gallery

Examples

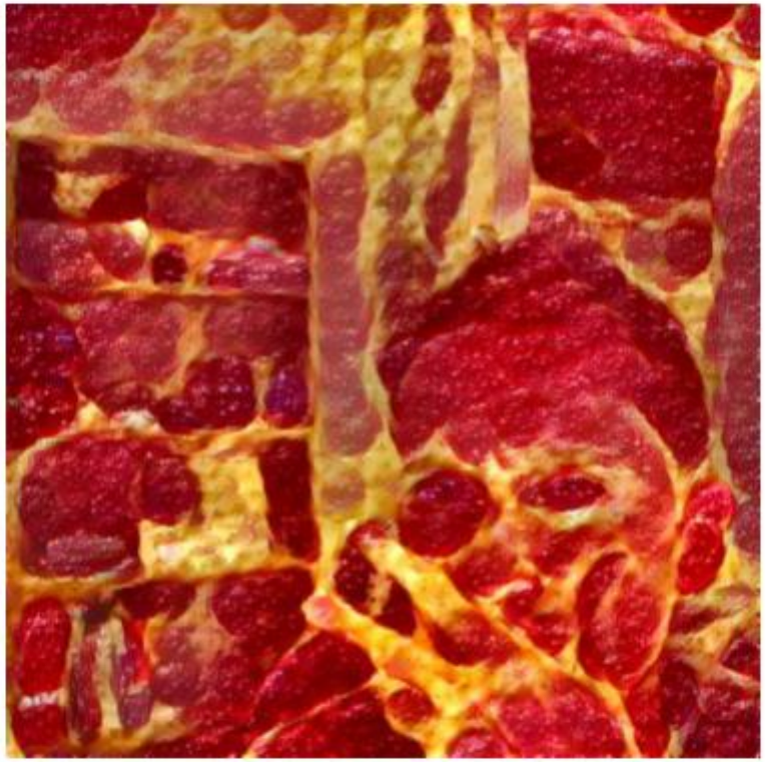
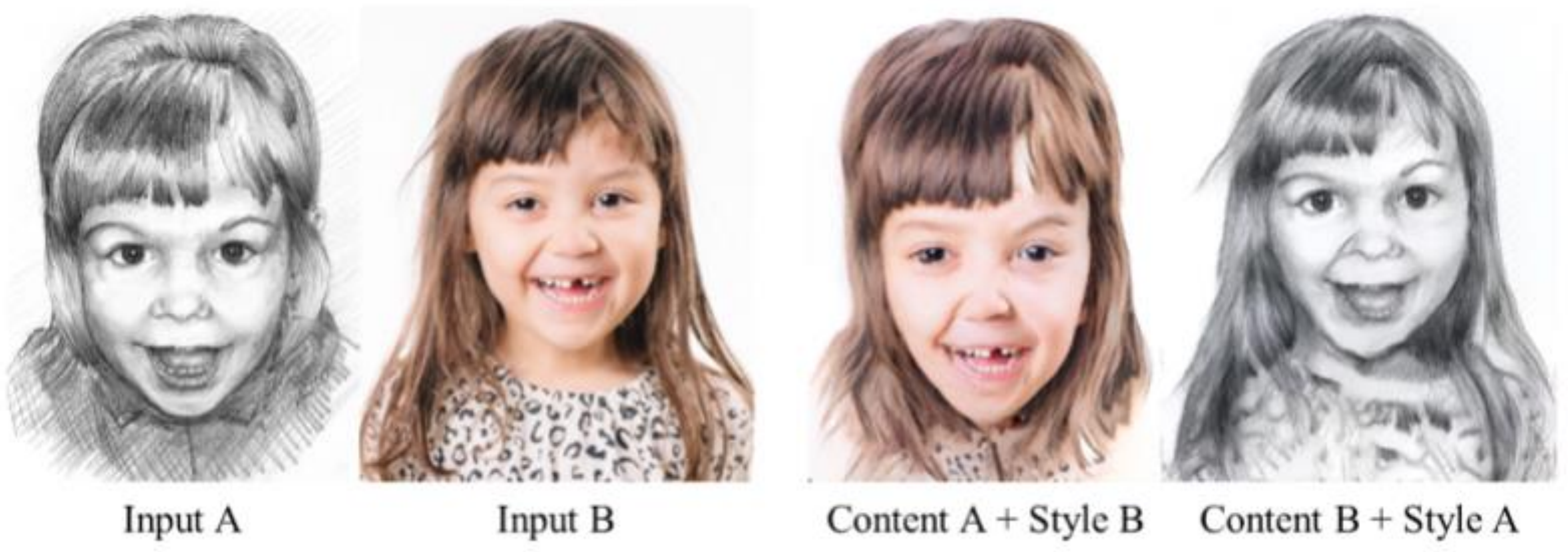


Figure: **Left:** My friend Grant, **Right:** Grant as a pizza

Artistic Style Transfer

- Recent methods combine CNNs with graphical models (CPSC 540):



https://arxiv.org/pdf/1601.04589.pdf

Artistic Style Transfer

- Recent methods combine CNNs with graphical models (CPSC 540):

https://arxiv.org/pdf/1601.04589.pdf

Artistic Style Transfer for Video

- Combining style transfer with optical flow:
  - https://www.youtube.com/watch?v=Khuj4ASldmU
- Videos from a former CPSC 340 student/TA's paper: