CPSC 340:
Machine Learning and Data Mining

More CNNs

Spring 2022 (2021W2)

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.

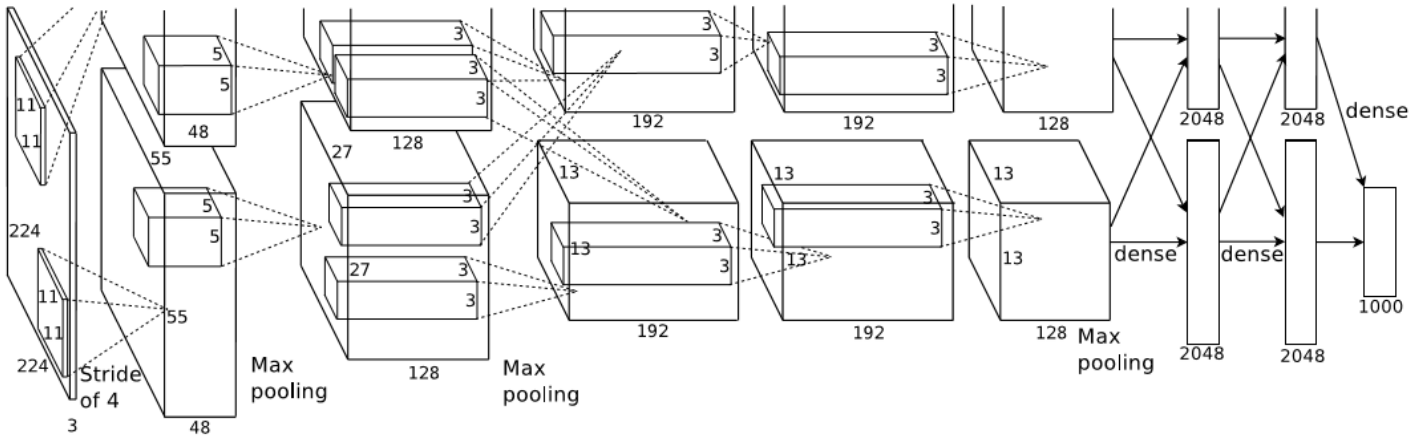
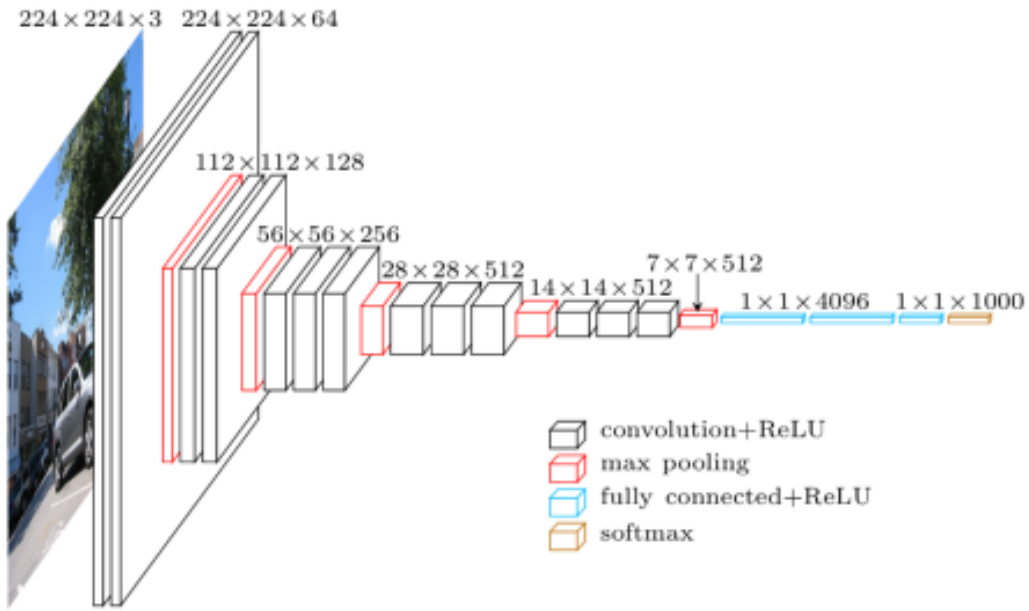


Figure 2: An illustration of the architecture of our CNN, explicitly showing the delineation of responsibilities between the two GPUs. One GPU runs the layer-parts at the top of the figure while the other runs the layer-parts at the bottom. The GPUs communicate only at certain layers. The network's input 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

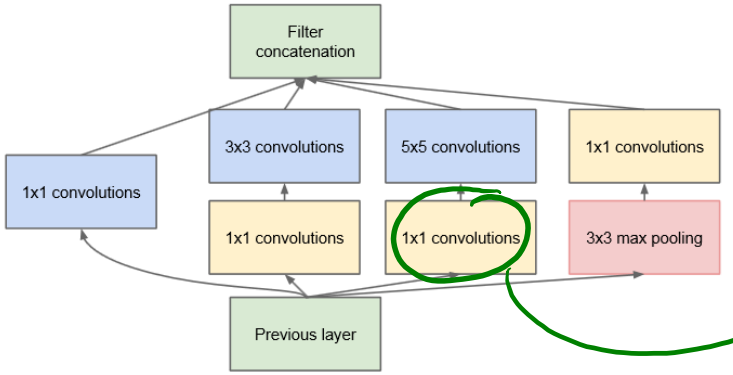
- Filters and stride got smaller over time.
  - Popular VGG approach uses 3x3 convolution layers with stride of 1.
    - 3x3 followed by 3x3 simulates a 5x5, and another 3x3 simulates a 7x7, and so on.
    - Speeds things up and reduces number of parameters.
    - Increases number of non-linear ReLU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



(b) Inception module with dimensionality reduction

"1 x 1" 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.

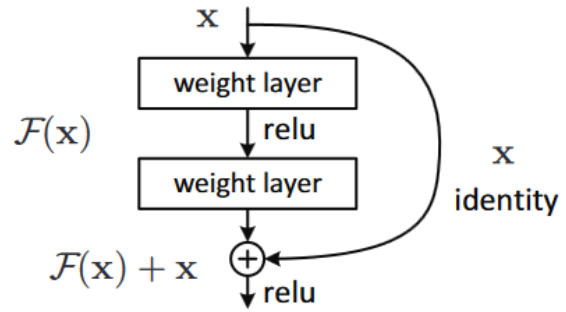


Figure 2. Residual learning: a building block.

Are CNNs learning something sensible?

- Filters learned by first layer of original AlexNet:

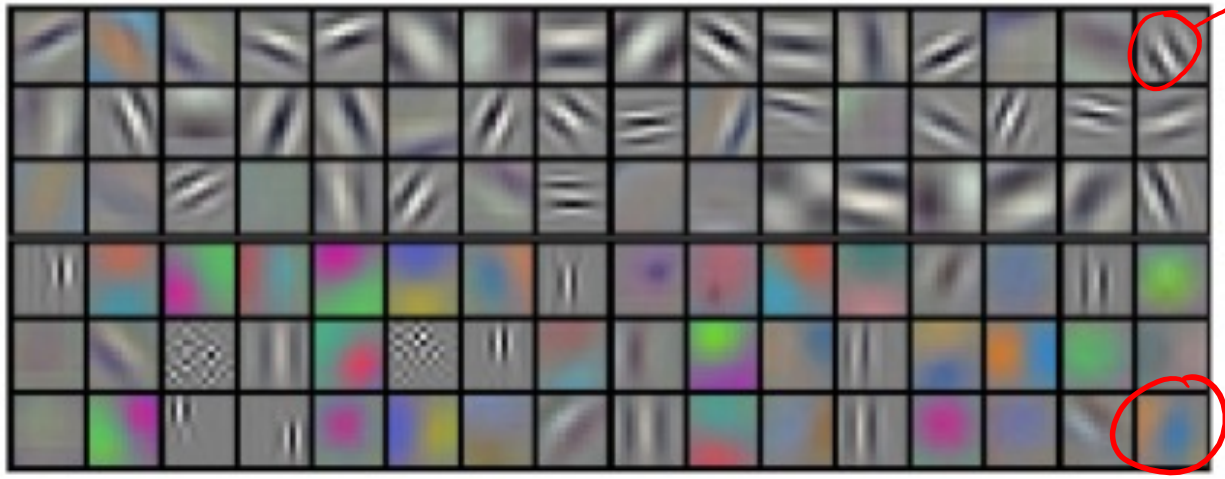


Figure 3: 96 convolutional kernels of size $11 \times 11 \times 3$ learned by the first convolutional layer on the $224 \times 224 \times 3$ input images. The

- Note that non-orthogonal PCA gives similar results (but 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.
  - Deconvolution networks try to reconstruct what "activates" filters.

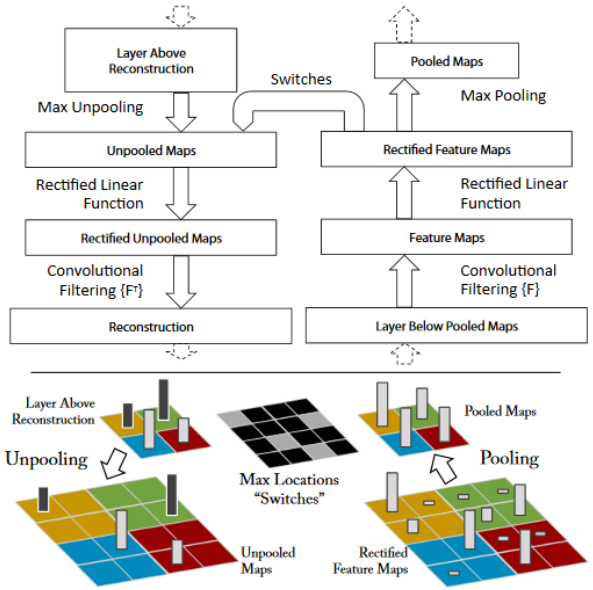
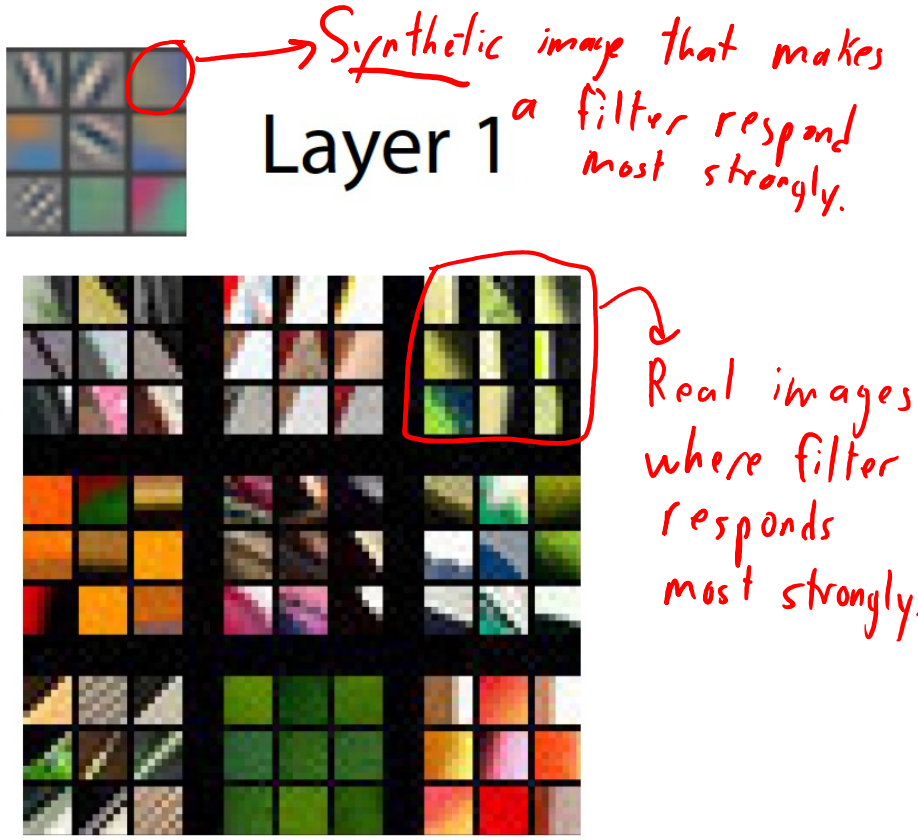
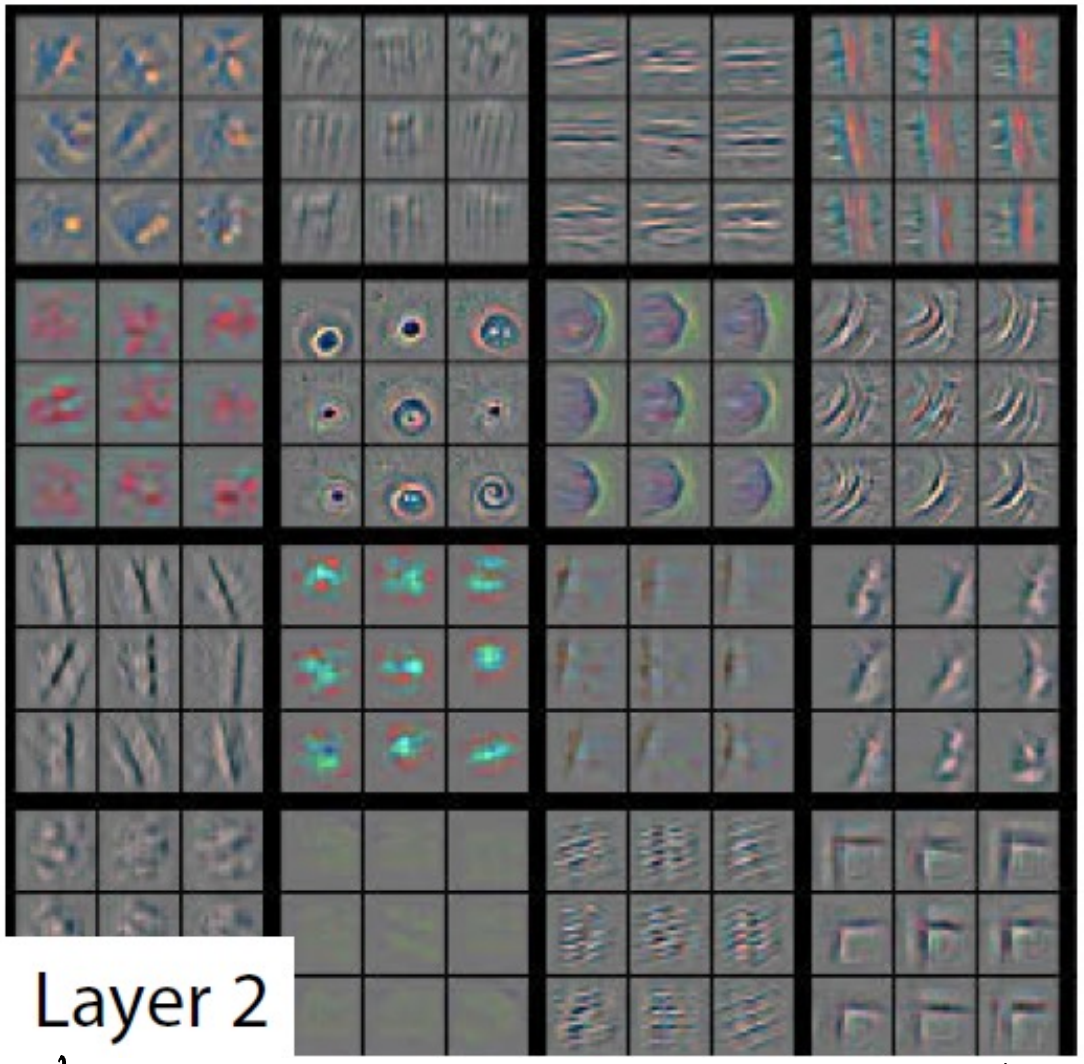


Figure 1. Top: A deconvnet layer (left) attached to a convnet layer (right). The deconvnet will reconstruct an approximate version of the convnet features from the layer beneath. Bottom: An illustration of the unpooling operation in the deconvnet, using *switches* which record the location of the local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

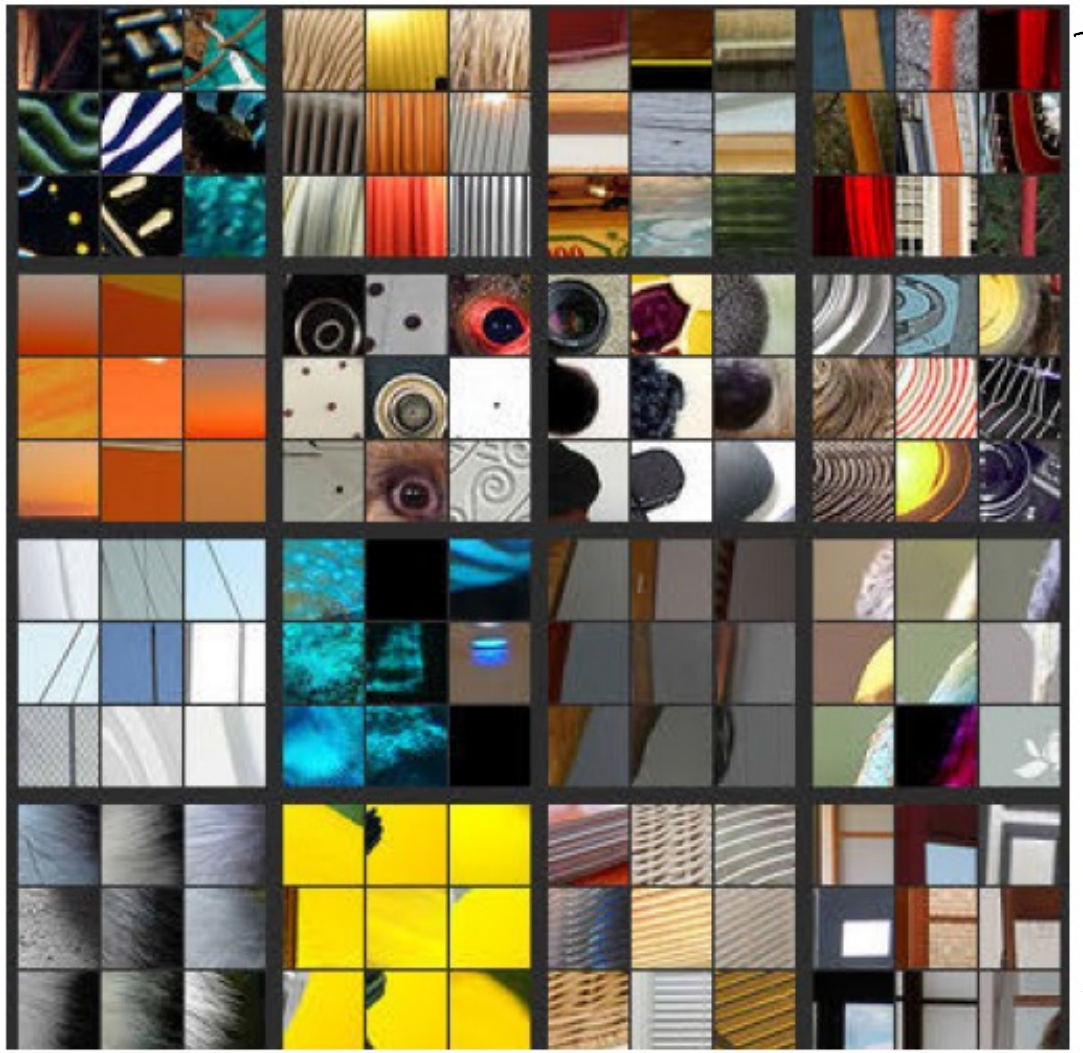


https://arxiv.org/pdf/1311.2901v3.pdf

Are CNNs learning something sensible?



Layer 2

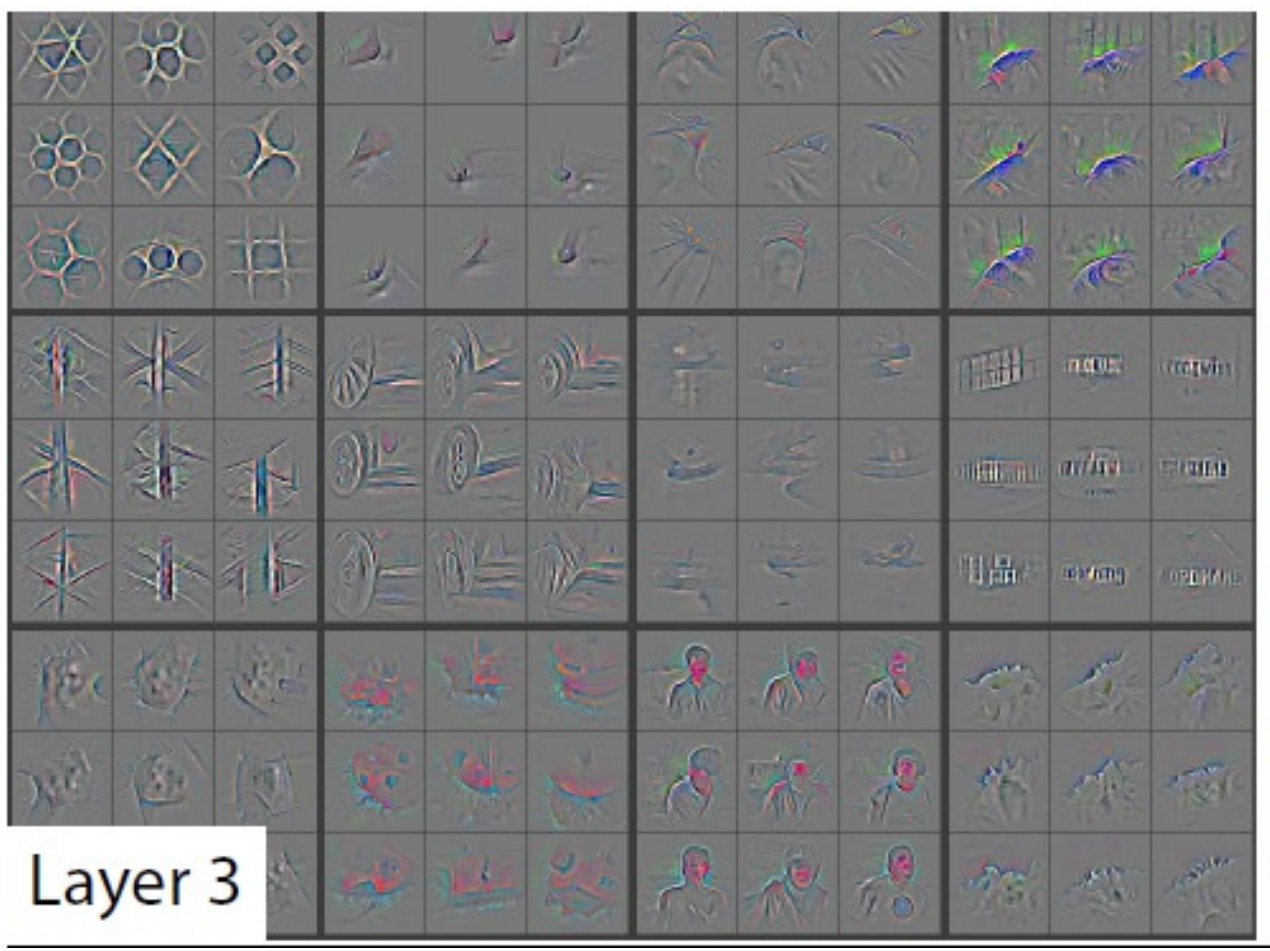


Patch from data giving largest response

Deconvolution network giving patch that leads to largest response

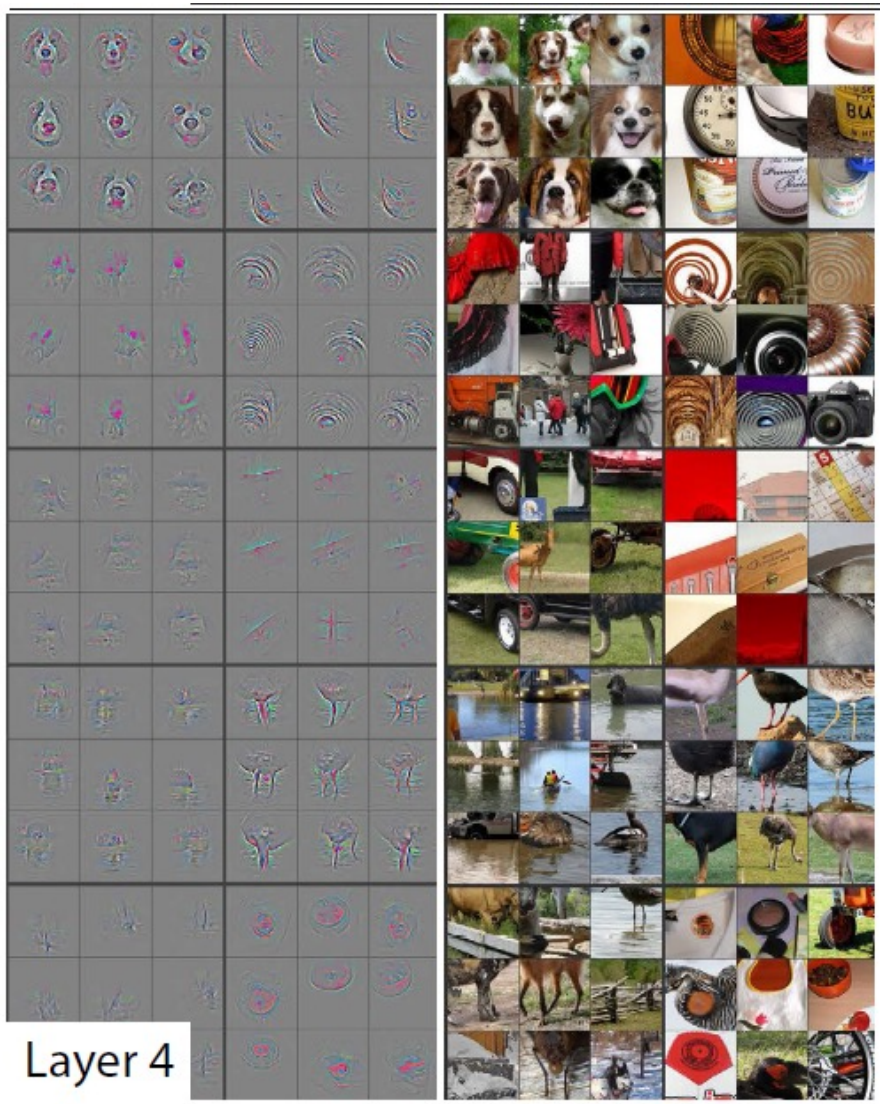
https://arxiv.org/pdf/1311.2901v3.pdf

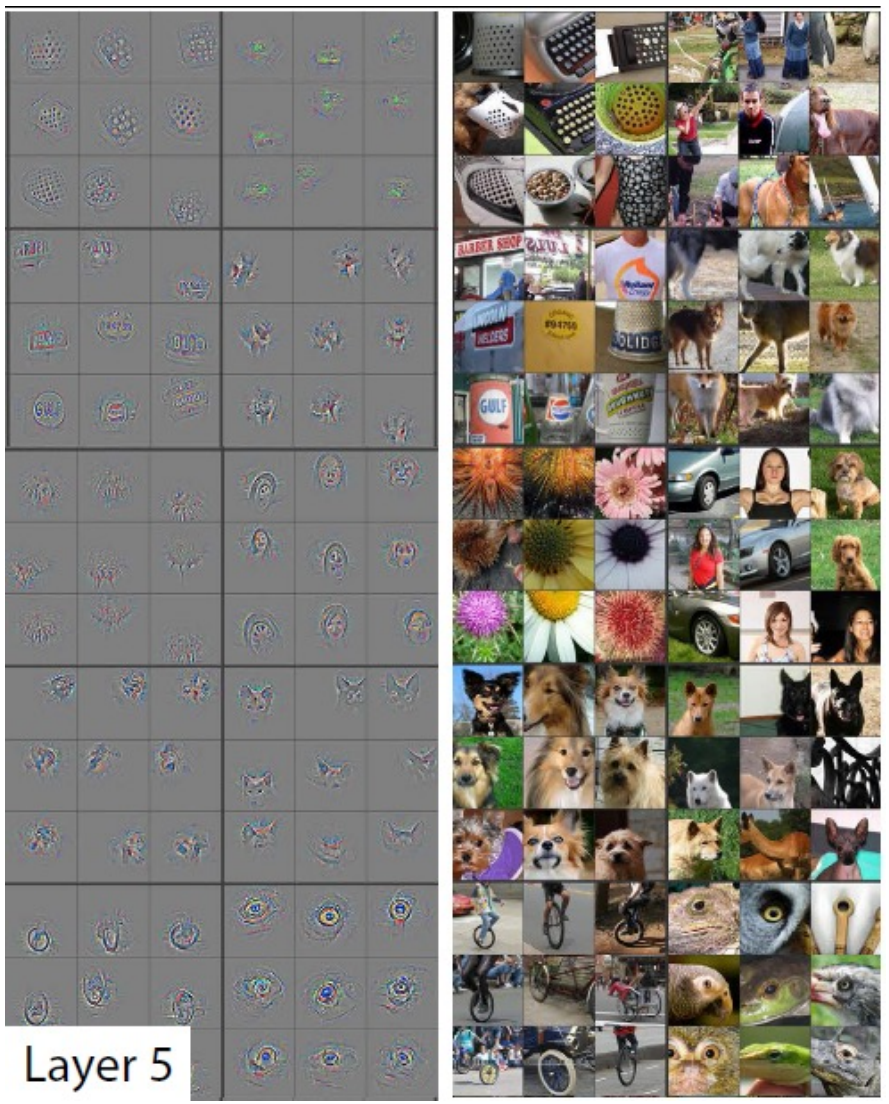
Are CNNs learning something sensible?



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

Are CNNs learning something sensible?

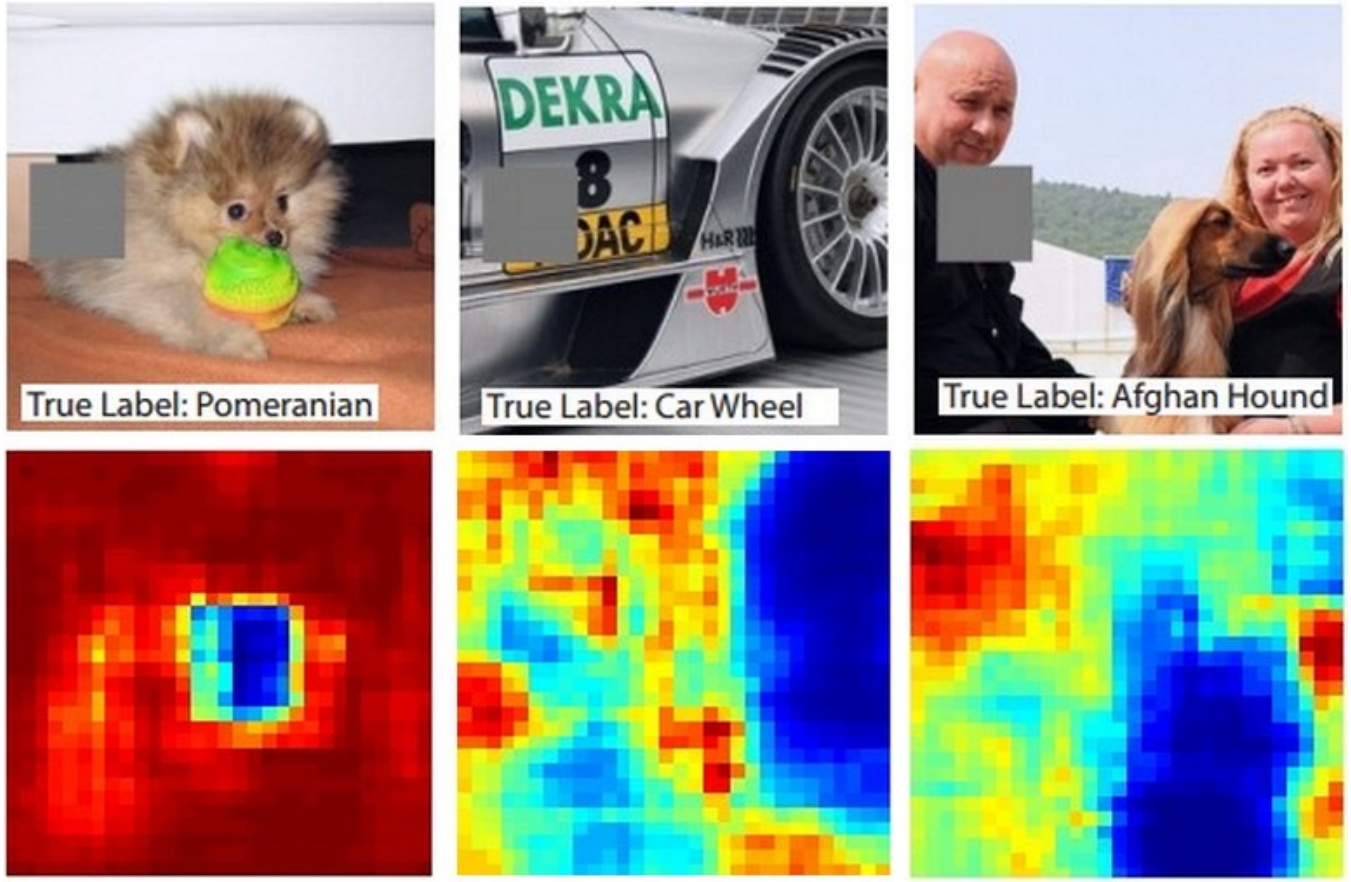




https://arxiv.org/pdf/1311.2901v3.pdf

Are CNNs learning something sensible?

- We can look at how output probability changes if we hide different parts of the input image:



<- predicted prob of true label if the occlusion is here

(we see low prob when the actual object is hidden)

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 maybe some overfitting to popular datasets like ImageNet.
    - Recent work showed accuracy drop of 10-15% with a different test set… but the ordering of models was almost unchanged.

- CNNs are now making their way into products.
  - Face recognition.
  - Amazon Go: https://www.youtube.com/watch?v=NrmMk1Myrxc
  - 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").*

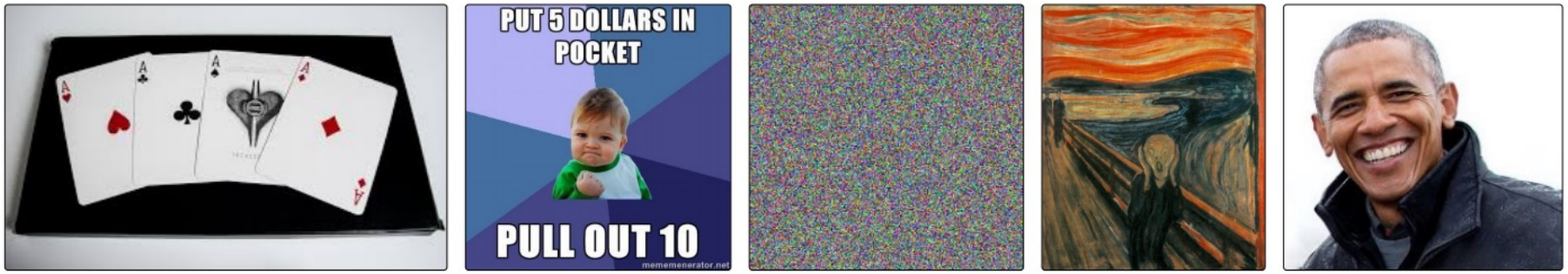
*Our training broke. Our error rate went from <25% error to ~99.97% error (on a standard 0-1 binary loss).

- "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 for the last few years.

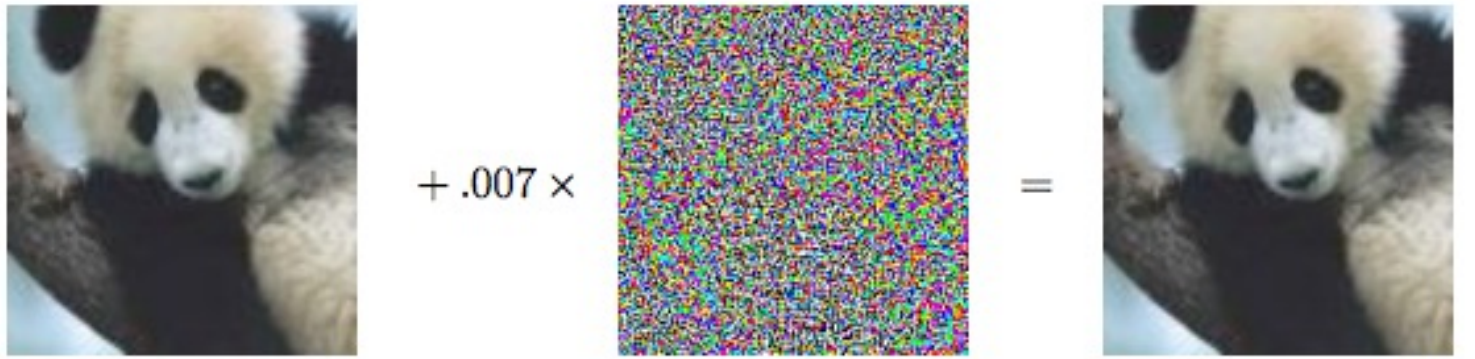


DenseNet 161 (2017)	Envelope 31%	Balance Beam 52%	Chainlink Fence 31%	Chest 37%	Tench 36%
SqueezeNet (2016)	Binder 43%	Balance Beam 18%	Poncho 32%	Jean 30%	Suit 21%
ResNet 152 (2015)	Envelope 40%	Pacifier 33%	Chain Mail 29%	Dust Cover 52%	Sweatshirt 25%
VGG 19 (2014)	Binder 51%	Dust Cover 44%	Window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

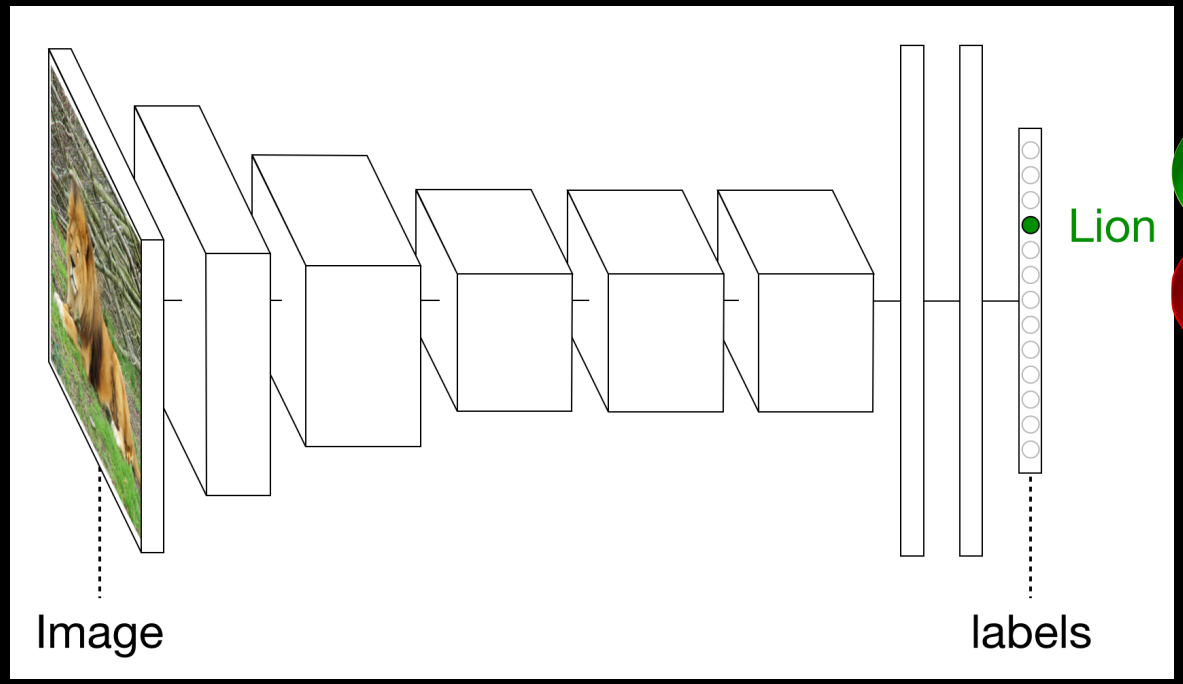
Mission Accomplished?

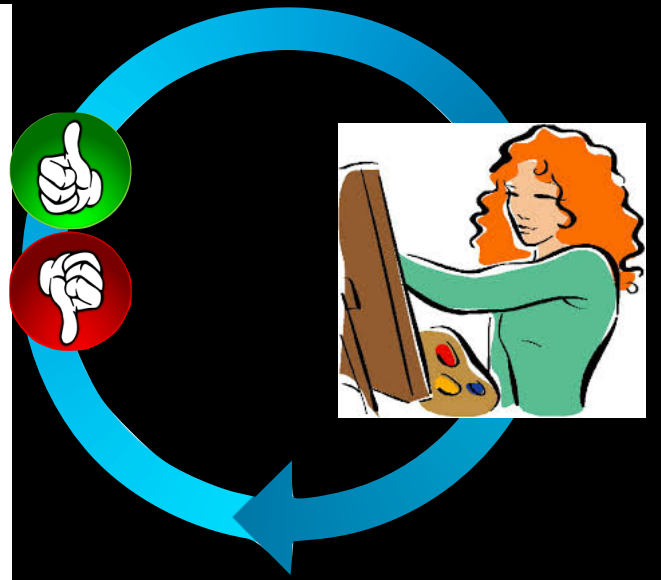
- Despite high-level of abstraction, deep CNNs are easily fooled:
  - Hot research topic at the moment.
- Recent work: imperceptible noise that changes the predicted label.
  - "Adversarial" examples (can change to any other label).



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

Investigating What Each Neuron Does

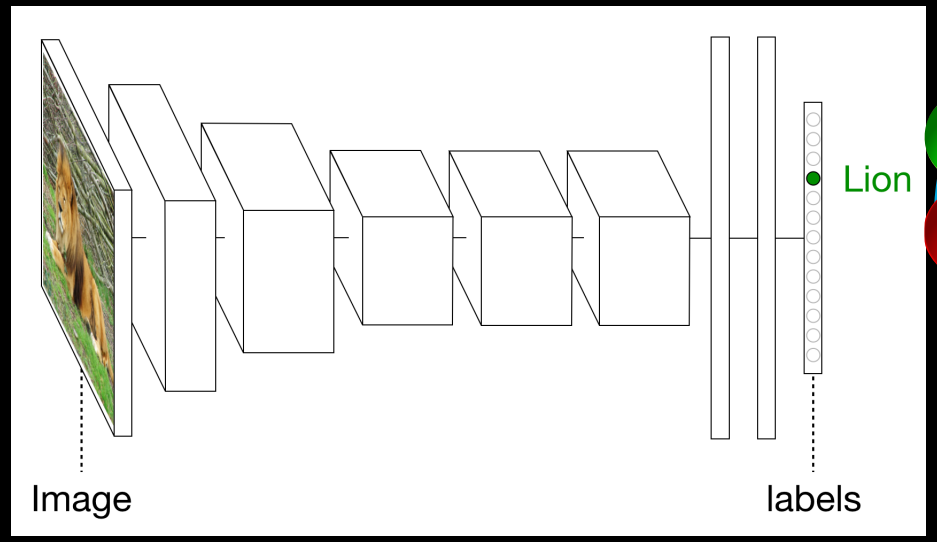


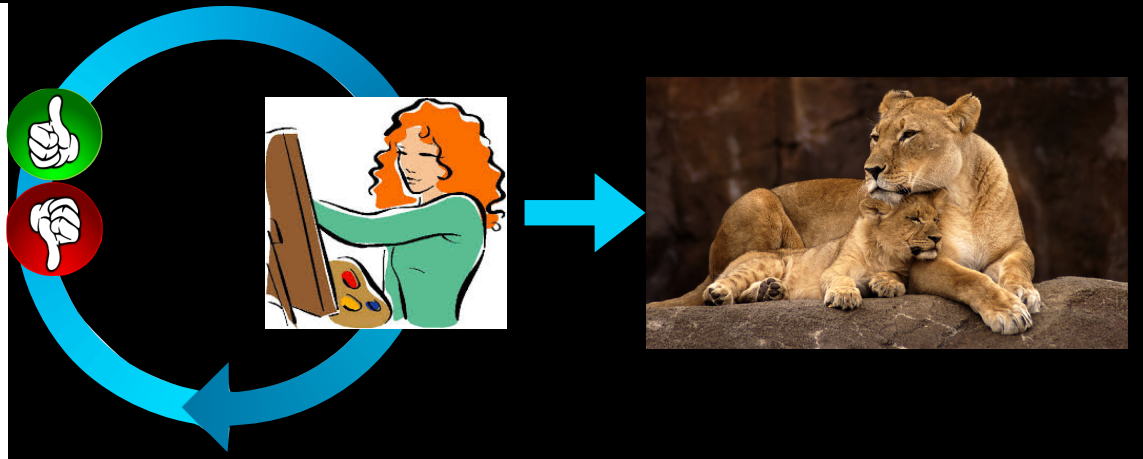


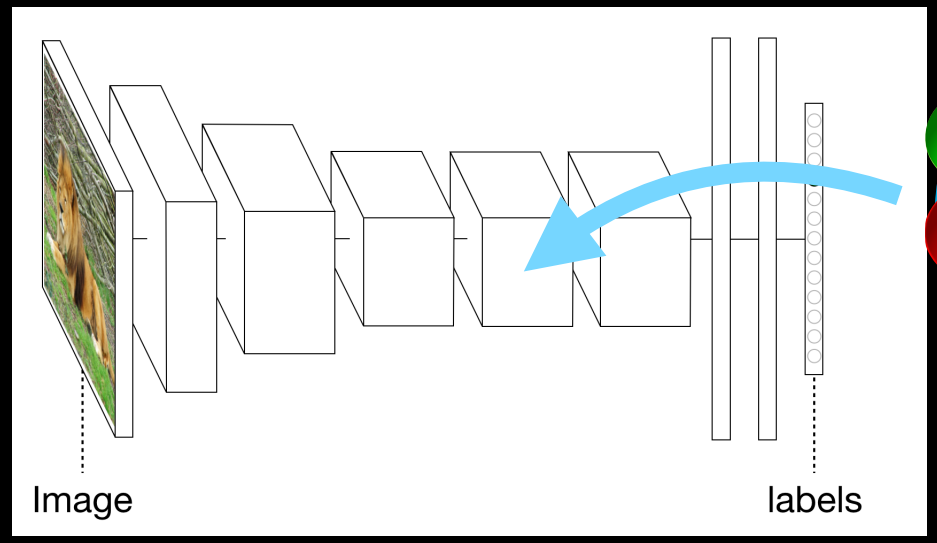
Pretrained, Fixed DNN

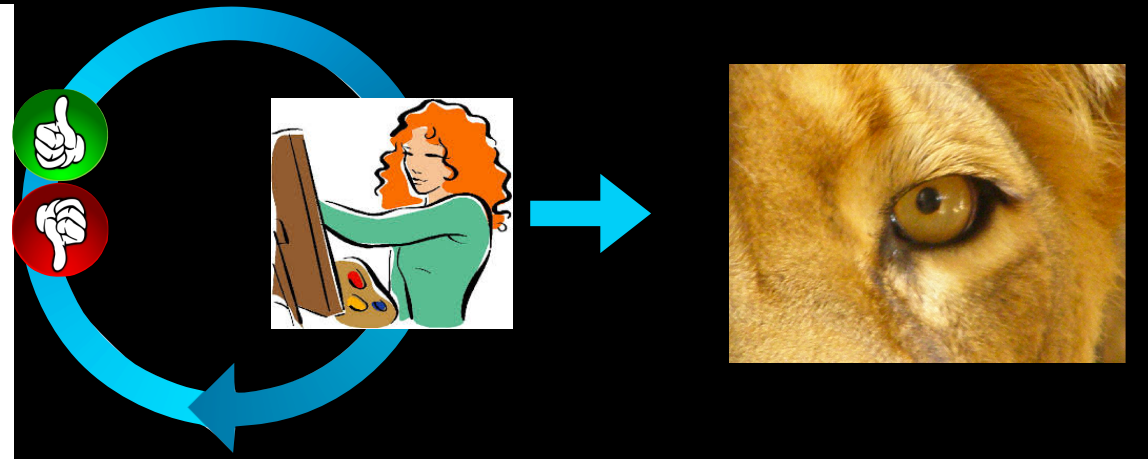
Optimize Pixels
e.g. via Backprop

“Deep Visualization”



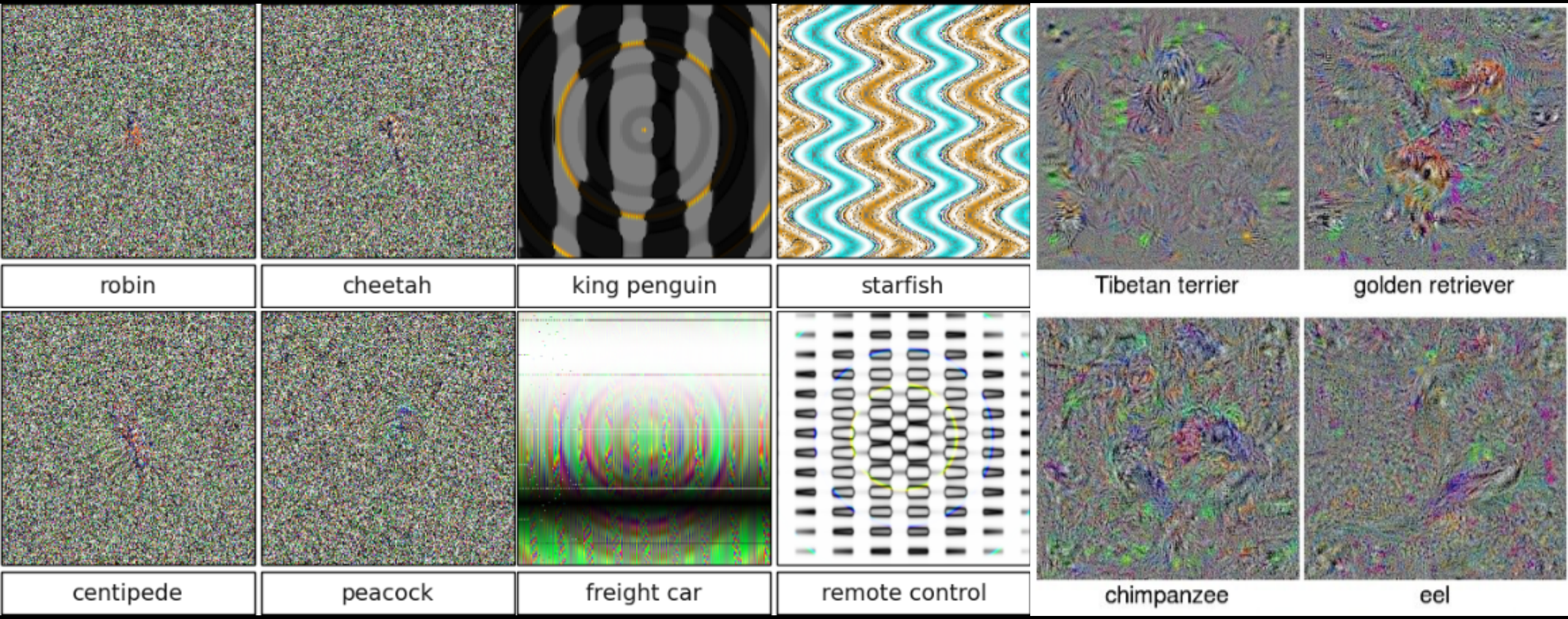






Deep Visualization Take 1

Nguyen, Yosinski, Clune, 2015, CVPR



DNN Confidence: > 99.6 % for all

Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images

Anh Nguyen
University of Wyoming
anguyen8@uwyo.edu

Jason Yosinski
Cornell University
yosinski@cs.cornell.edu

Jeff Clune
University of Wyoming
jeffclune@uwyo.edu

Abstract

Deep neural networks (DNNs) have recently been achieving state-of-the-art performance on a variety of pattern-recognition tasks, most notably visual classification problems. Given that DNNs are now able to classify objects in images with near-human-level performance, questions naturally arise as to what differences remain between computer and human vision. A recent study [30] revealed that changing an image (e.g. of a lion) in a way imperceptible to humans can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library). Here we show a related result: 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 (e.g. labeling with certainty that white noise static is a lion). Specifically, we take convolutional neural networks trained to perform well on either the ImageNet or MNIST datasets and then find images with evolutionary algorithms or gradient ascent that DNNs label with high confidence as belonging to each dataset class. It is possible to produce images totally unrecognizable to human eyes that DNNs believe with near certainty are familiar objects, which we call "fooling images" (more generally, fooling examples). Our results shed light on interesting differences between human vision and current DNNs, and raise questions about the generality of DNN computer vision.

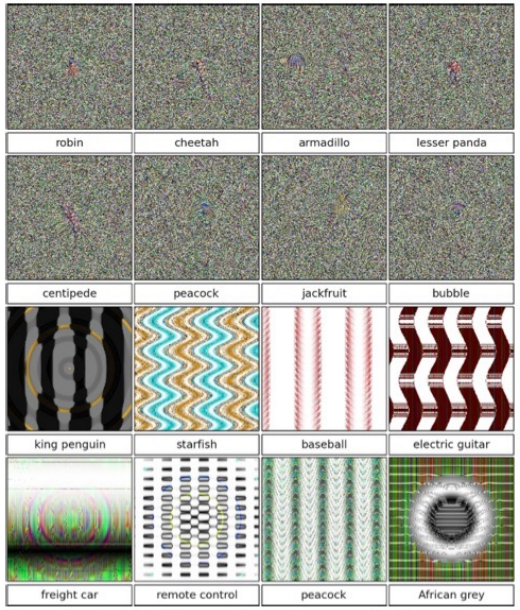


Figure 1. Evolved images that are unrecognizable to humans, but that state-of-the-art DNNs trained on ImageNet believe with $\geq$ 99.6% certainty to be a familiar object. This result highlights differences between how DNNs and humans recognize objects. Images are either directly (*top*) or indirectly (*bottom*) encoded.

1. Introduction

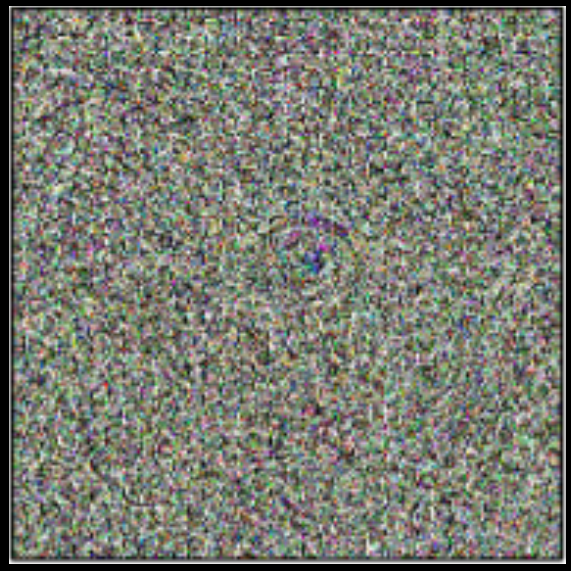
Deep neural networks (DNNs) learn hierarchical layers of representation from sensory input in order to perform pattern recognition [2, 14]. Recently, these deep architectures have demonstrated impressive, state-of-the-art, and sometimes human-competitive results on many pattern recognition tasks, especially vision classification problems [16, 7, 31, 17]. Given the near-human ability of DNNs to classify visual objects, questions arise as to what differences remain between computer and human vision.

A recent study revealed a major difference between DNN and human vision [30]. Changing an image, originally correctly classified (e.g. as a lion), in a way imperceptible to human eyes, can cause a DNN to label the image as something else entirely (e.g. mislabeling a lion a library).

In this paper, we show another way that DNN and human vision differ: It is easy to produce images that are completely unrecognizable to humans (Fig. 1), but that state-of-the-art DNNs believe to be recognizable objects with over 99% confidence (e.g. labeling with certainty that TV static

- May not understand much
- Huge security concern
- Helped launch avalanche of work into "adversarial & fooling examples"
  - with Szegedy et al. 2013

School bus



Open road!

Why are networks easily fooled?

https://www.youtube.com/watch?v=3lp9eN5JE2A

Mission Accomplished?

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

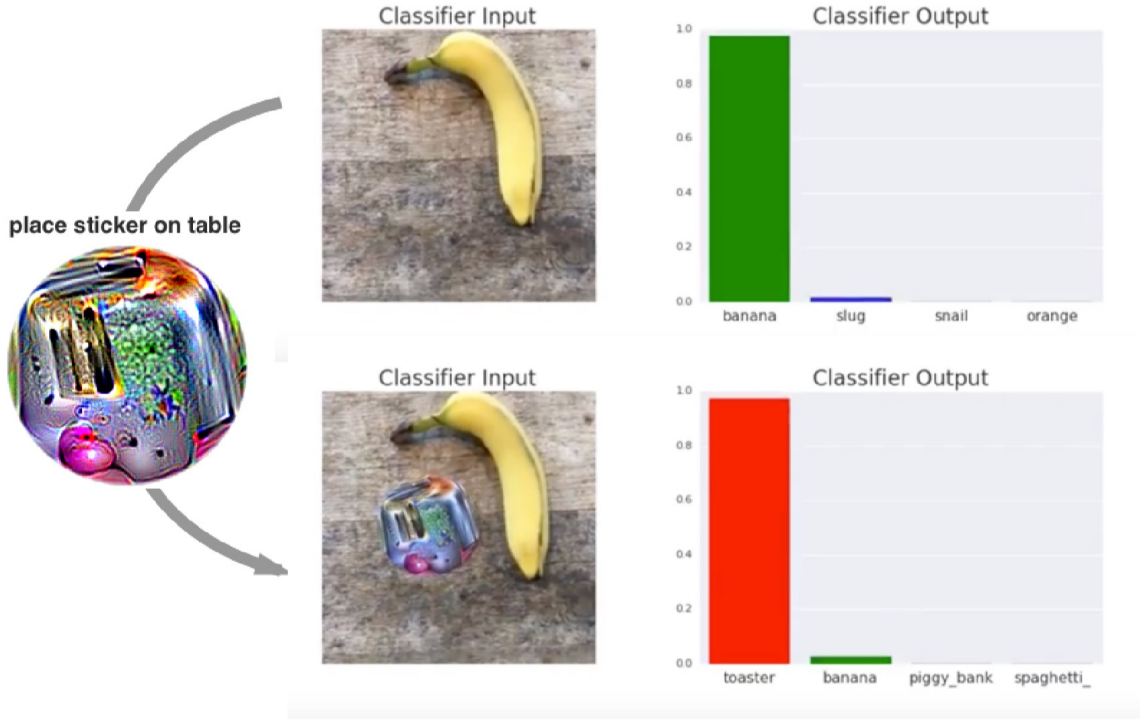


Figure 1: A real-world attack on VGG16, using a physical patch generated by the white-box ensemble method described in Section 3. When a photo of a tabletop with a banana and a notebook (top photograph) is passed through VGG16, the network reports class 'banana' with 97% confidence (top plot). If we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

Eykholt et al. 2018

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

Mission Accomplished?

- …or can it be even easier?

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

Granny Smith	0.3%
iPod	0.0%
library	0.0%
pizza	0.0%
toaster	93.7%
dough	0.2%

https://openai.com/blog/multimodal-neurons/

Mission Accomplished?

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

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

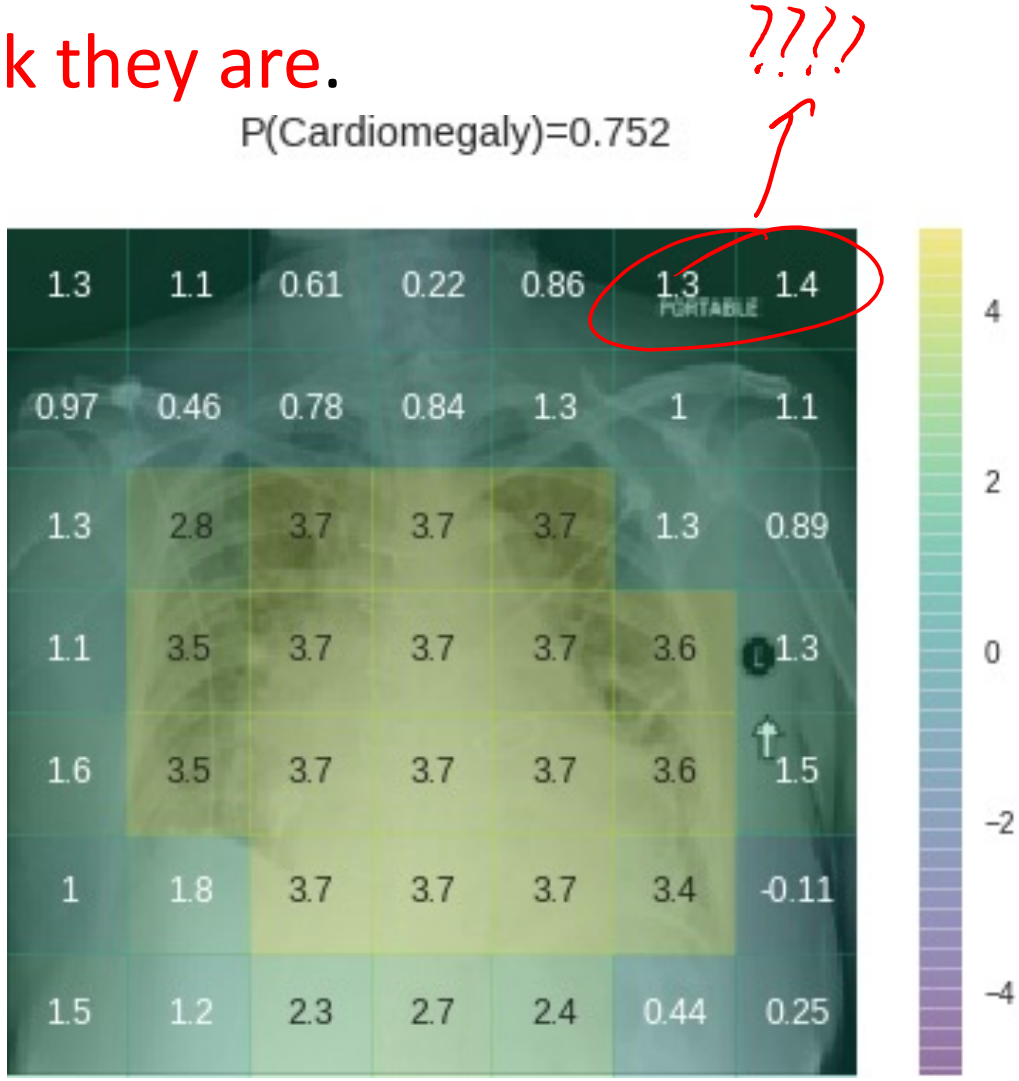
(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

Beery / van Horn / Perona, “Recognition in Terra Incognita”, ECCV 2018, based on ClarifAI.com API

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” protocol is predictive:
    - But they are probably getting a “portable” scan because they’re too sick to go the hospital.
  - CNN was biased by the scanning protocol.
    - 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 – can be exacerbated by choice of regularizer / loss function).

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

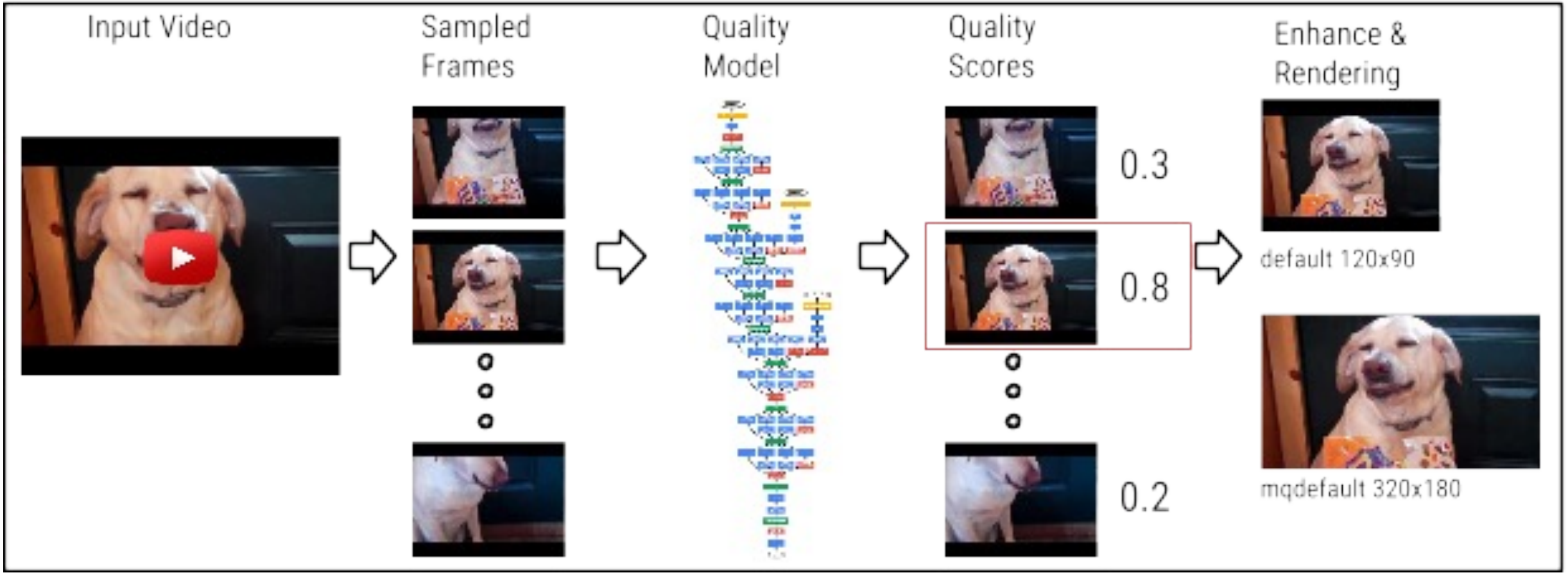
Energy Costs

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

- 2019 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.
  - But see counter (or mitigating) arguments here

(pause)

CNNs for Choosing YouTube Thumbnails



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

Beyond Classification (CPSC 440)

- “Fully convolutional” neural networks allow “dense” prediction:

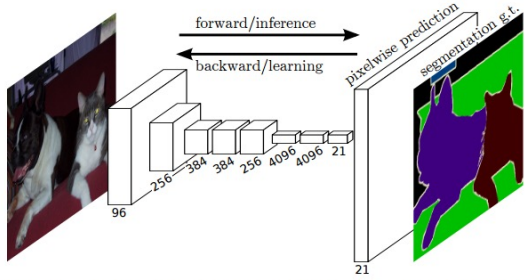


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

- Image segmentation:

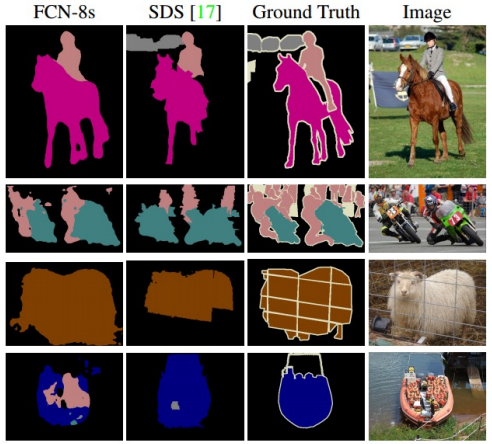
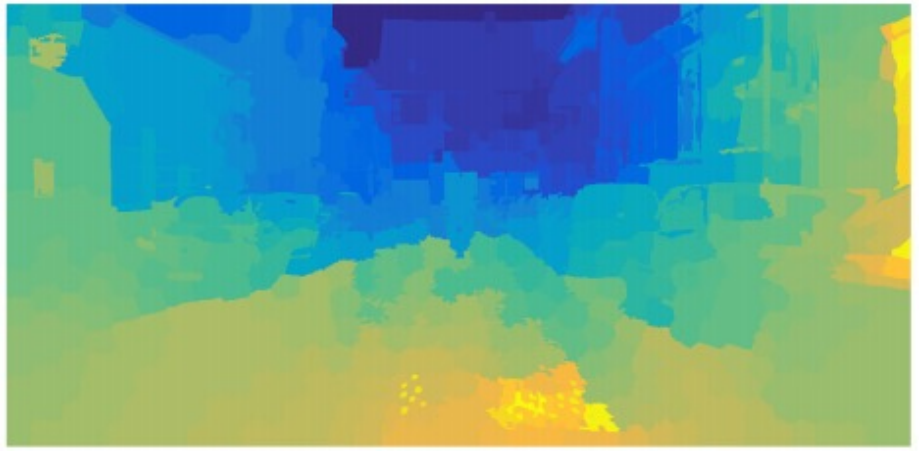


Figure 6. Fully convolutional segmentation nets produce state-of-the-art performance on PASCAL. The left column shows the output of our highest performing net, FCN-8s. The second shows the segmentations produced by the previous state-of-the-art system by Hariharan *et al.* [17]. Notice the fine structures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

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

Beyond Classification (CPSC 440)

- Depth Estimation:



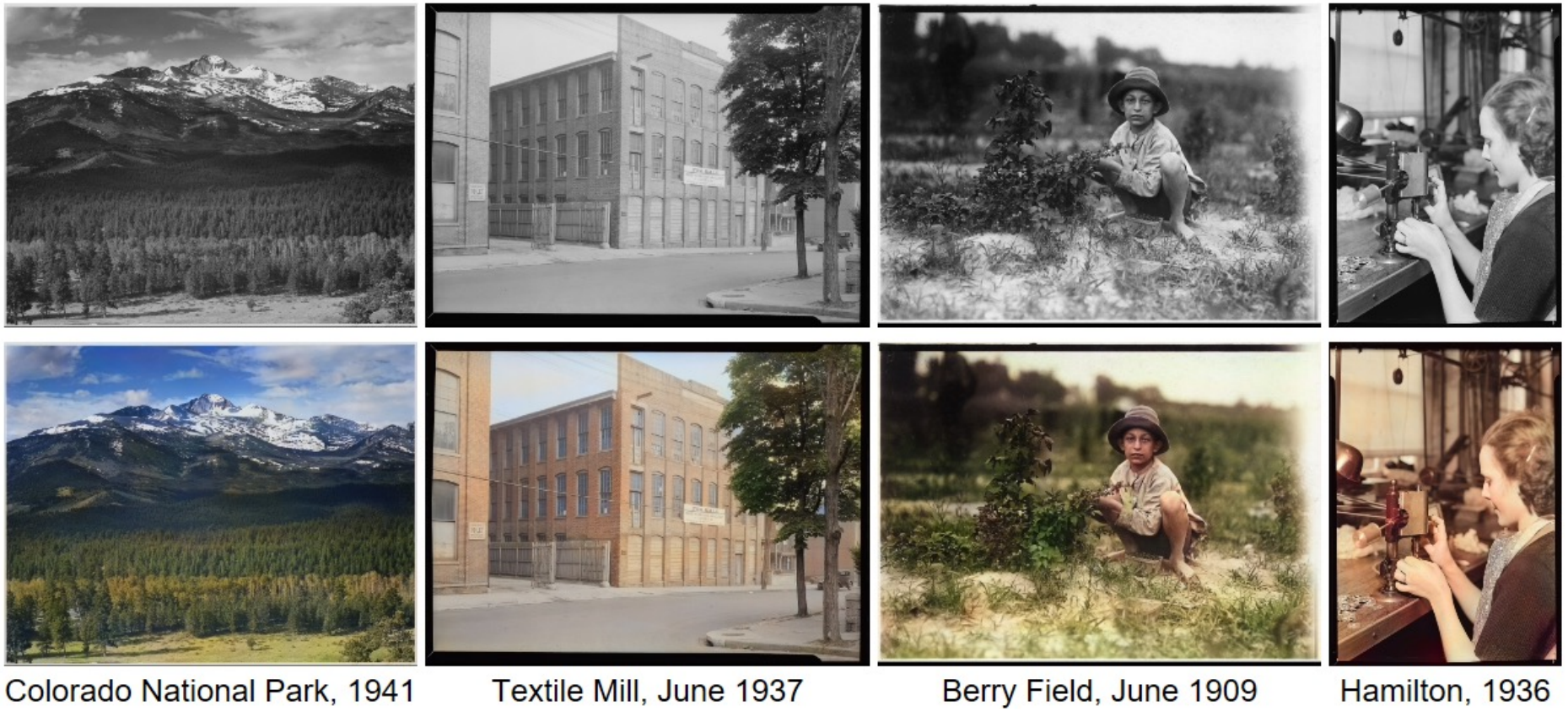
- "A Year in Computer Vision" from 2017 indicates how much progress happens each year

Beyond Classification (CPSC 440)

- “AutoPortrait”: automatic photo re-touching.

Beyond Classification (CPSC 440)

- Image colorization:



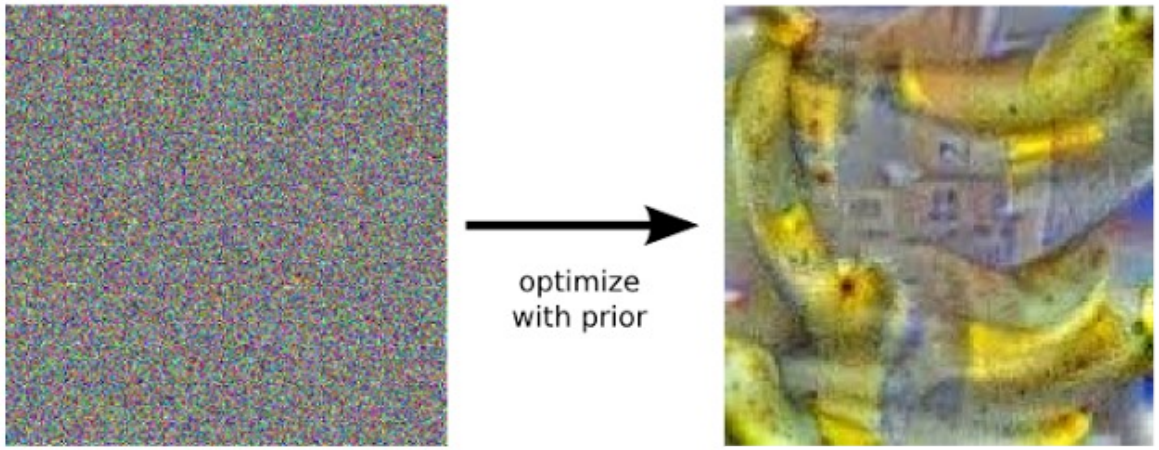
  - Image Gallery, Video

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

“Inceptionism” / Deep Dream

- Instead of choosing best weights, choose best input by running gradient descent on 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” / Deep Dream

- Inceptionism for different class labels:


Hartebeest


Measuring Cup


Ant


Starfish


Anemone Fish


Banana


Parachute


Screw

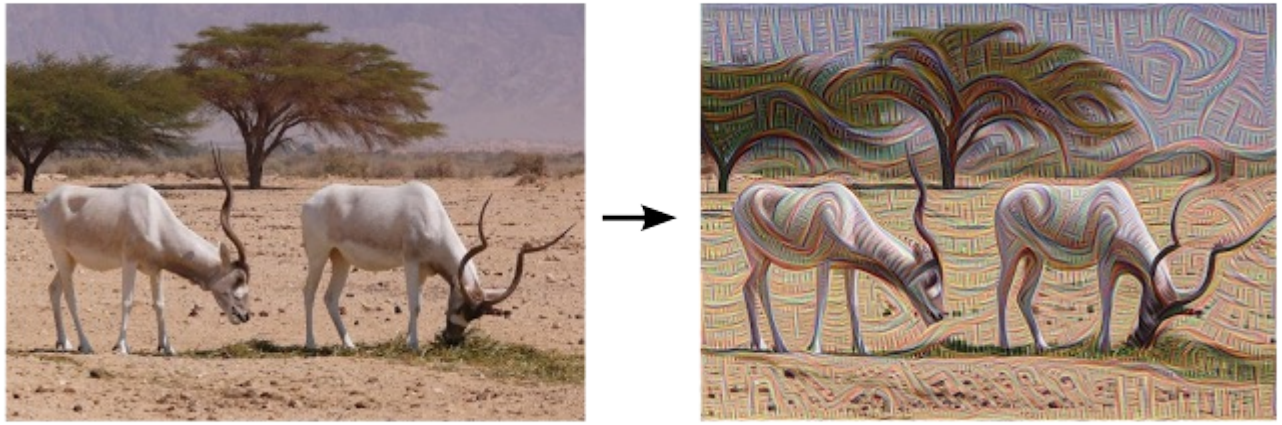
Dumbbell




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

“Inceptionism” / Deep Dream

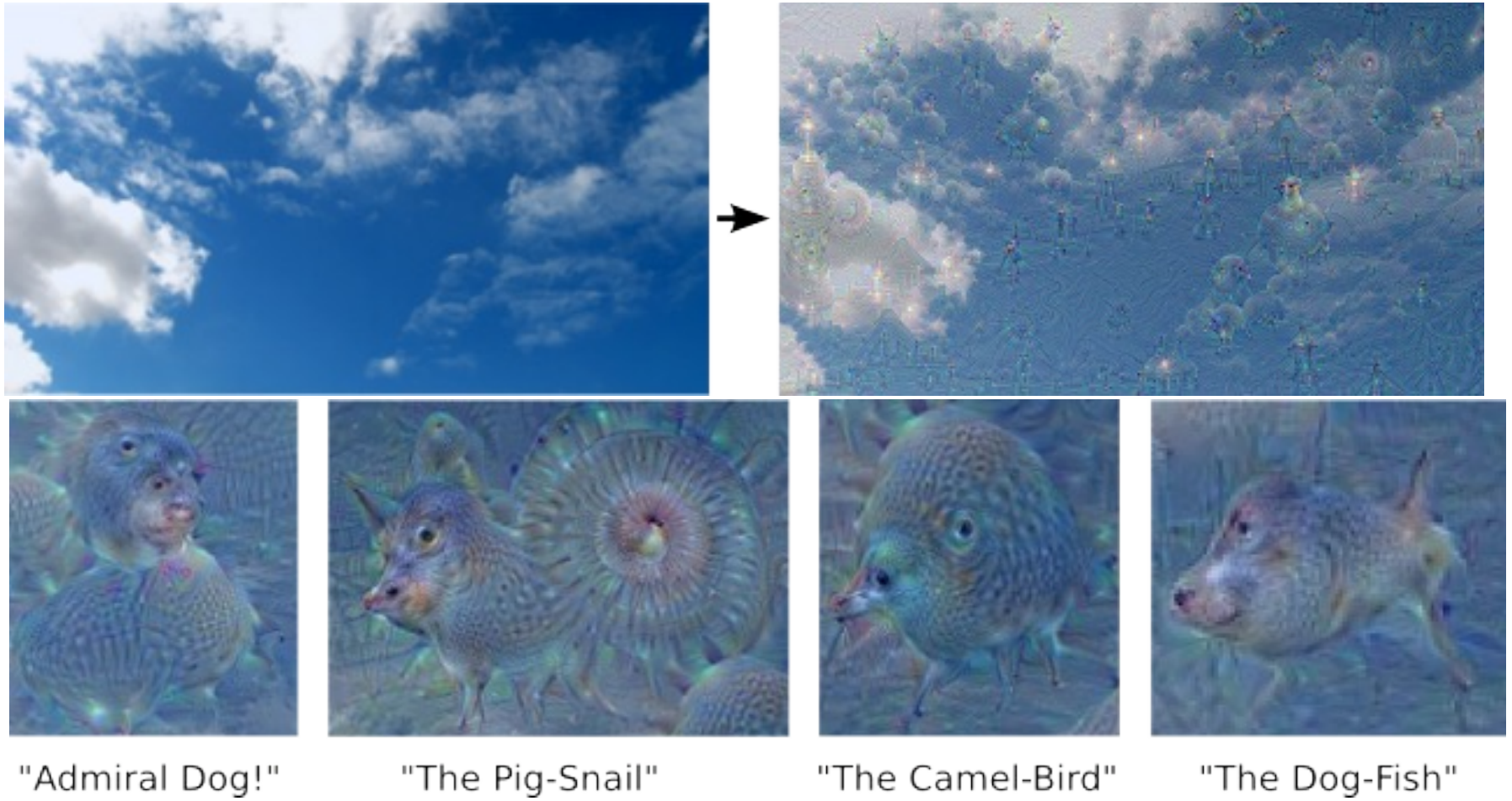
- Inceptionism where we try to match $z_i^{(m)}$ values instead of y_i.
  - Shallow ‘m’:



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

“Inceptionism” / Deep Dream

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



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

“Inceptionism” / Deep Dream

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

  - Deep Dream video

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’.

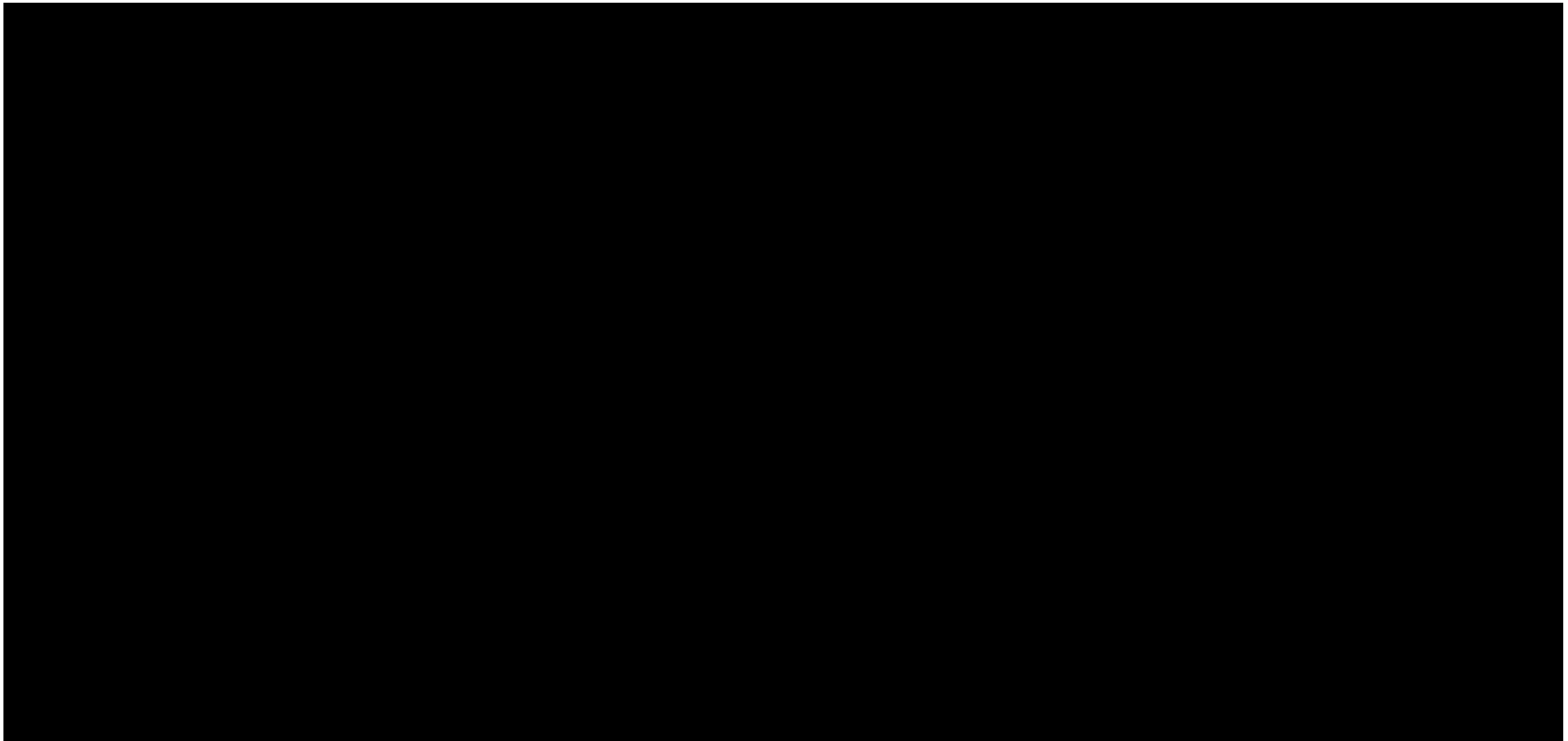
Artistic Style Transfer

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

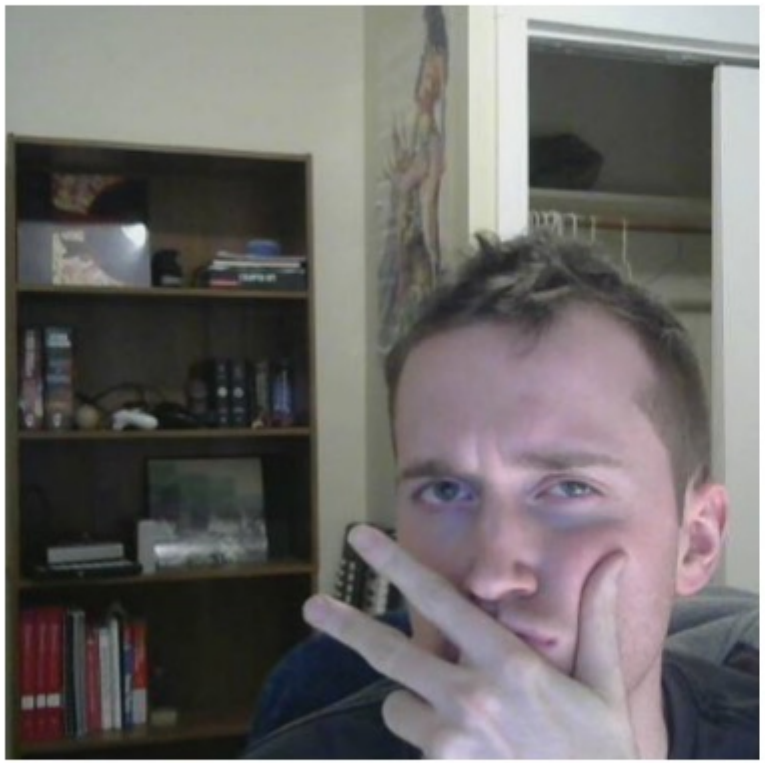
Image Gallery

Stable Style Transfer for Video

- https://www.youtube.com/watch?v=Khuj4ASldmU&t=17s



Examples



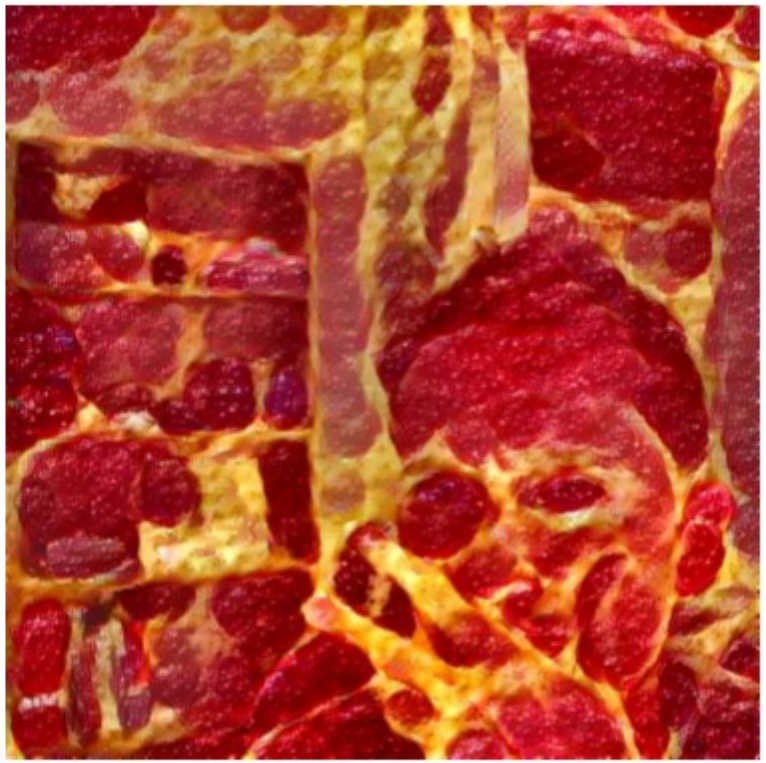


Figure: **Left:** My friend Grant, **Right:** Grant as a pizza

Artistic Style Transfer

- Recent methods combine CNNs with graphical models (CPSC 440):

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

Artistic Style Transfer

- Recent methods combine CNNs with graphical models (CPSC 440):

Input style

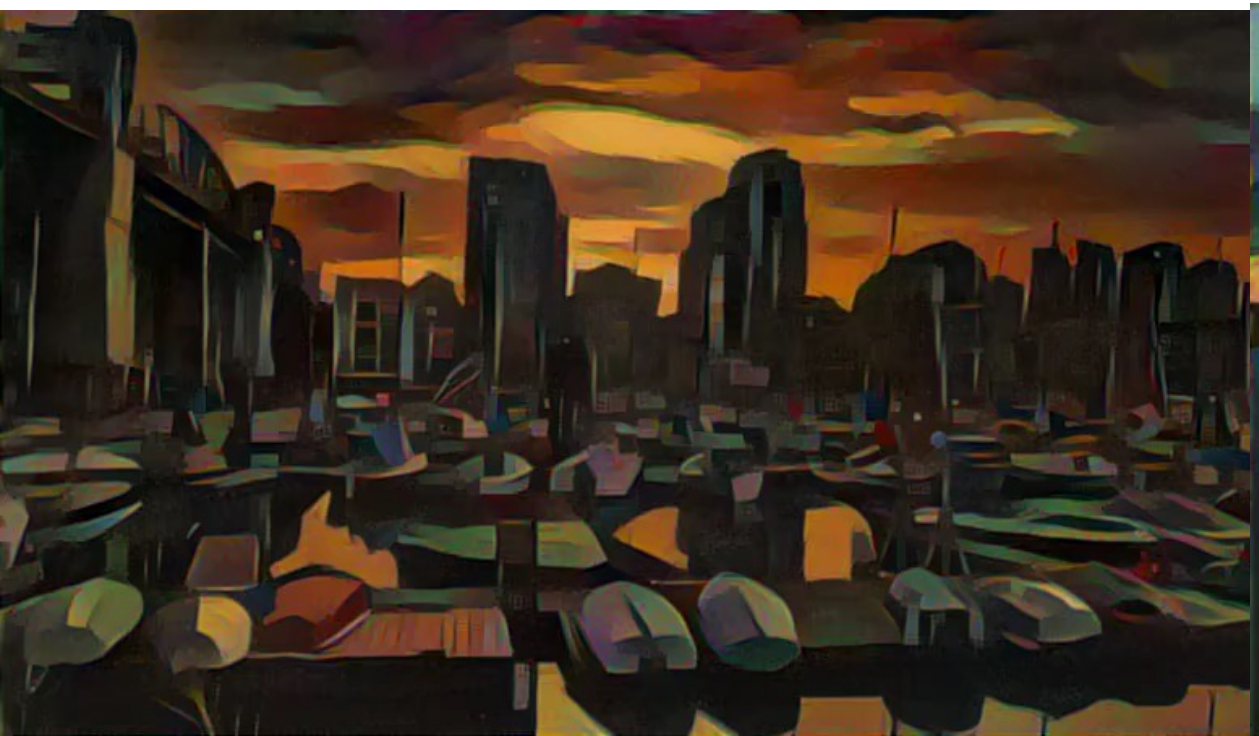
Input content

Ours

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:



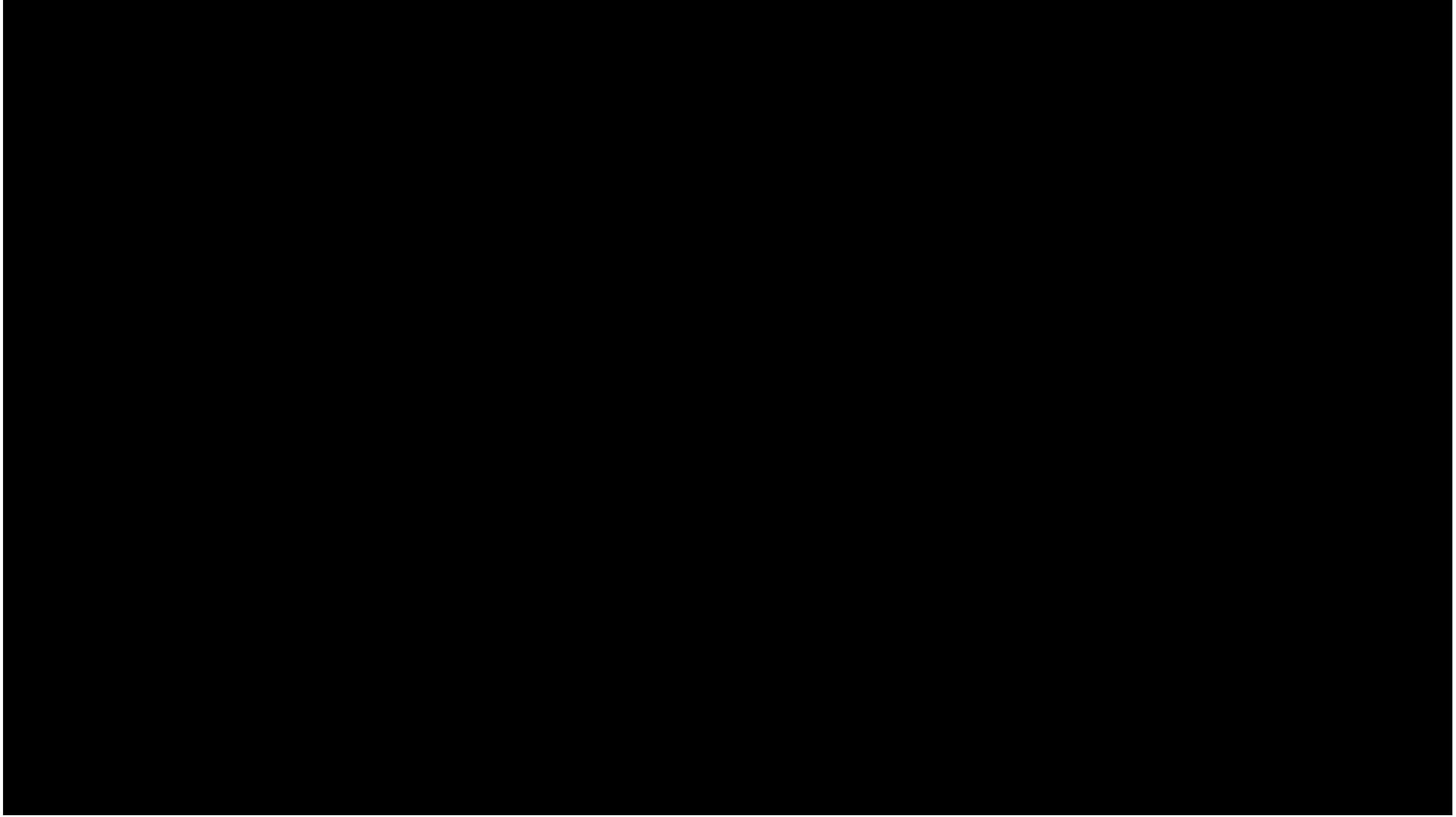
Generative Adversarial Networks (GANs)

GAN PROGRESS ON FACE GENERATION

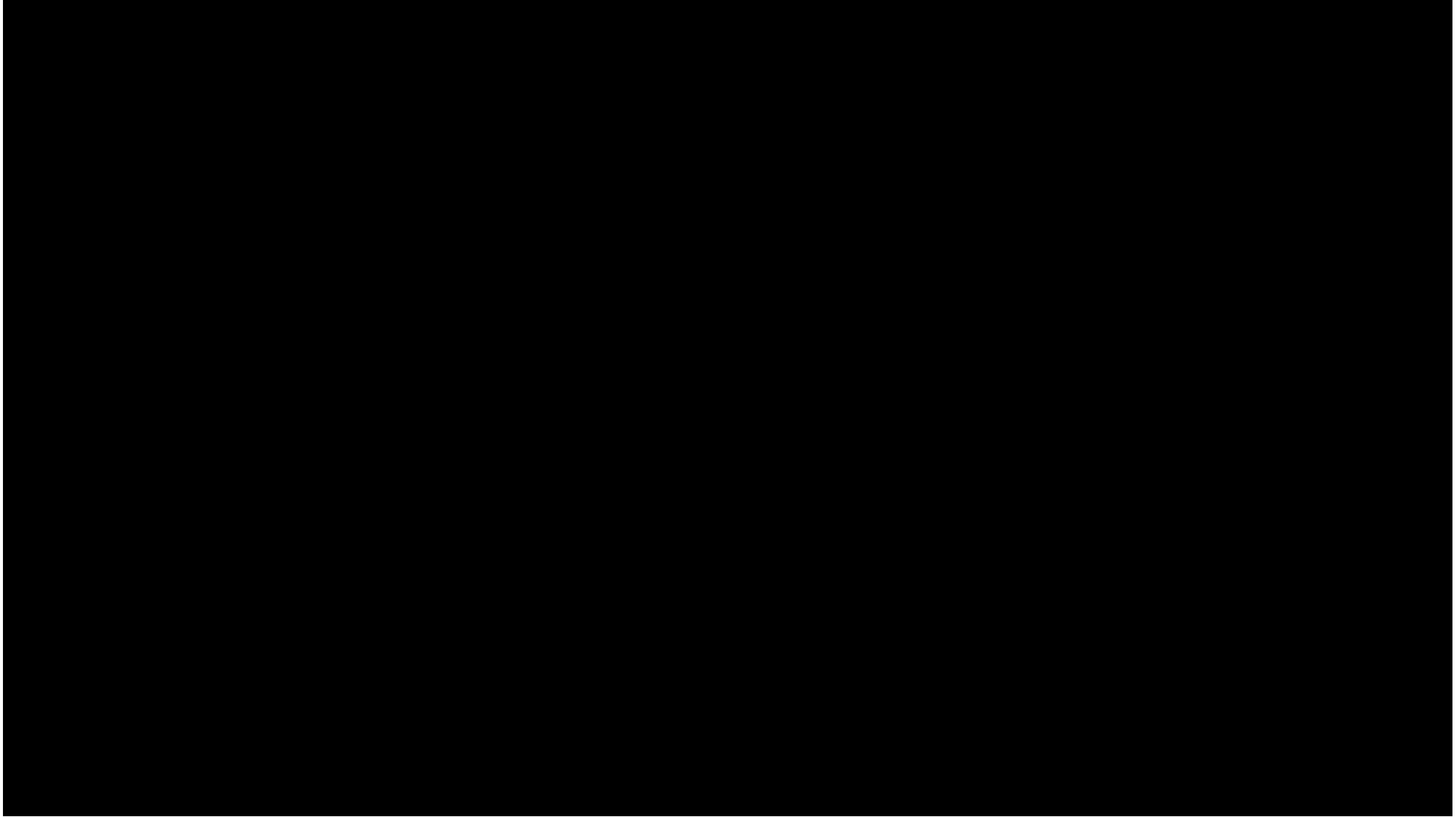
Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021

Figure 2.1.7

https://this-person-does-not-exist.com/en



2018



2022

DeepFakes

Top stories

PHYS.ORG

Deepfakes and fake news pose a growing threat to democracy, experts warn

1 hour ago

DB The Daily Beast

You Won't Believe What This 'Deepfake' Sean Hannity Did

1 day ago

CNN

Deepfakes are now trying to change the course of war

1 week ago

GIZMODO

Move Over Global Disinformation Campaigns, Deepfakes Have a New Rol...

6 days ago

Text→ Image

- Dall-e: https://openai.com/blog/dall-e

an armchair in the shape of an avocado. . . .

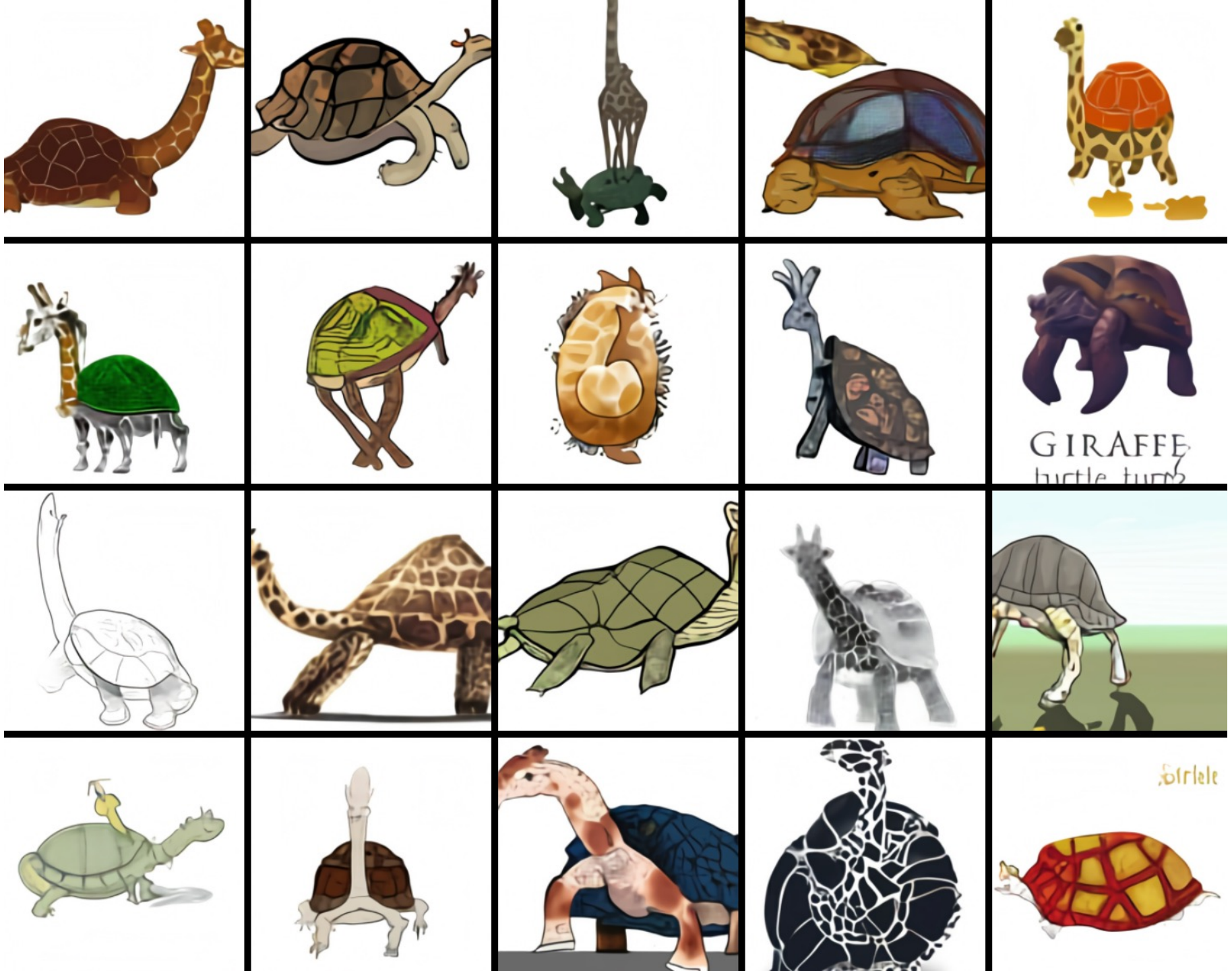
AI-GENERATED IMAGES

Edit prompt or view more images↓

a store front that has the word ‘openai’ written on it. . . .

AI-GENERATED IMAGES

Edit prompt or view more images↓


GIRAFFE

GPT-3

- Deep neural network
  - Transformer (one of the most important recent advances)
  - Generates next word
    - ~passes Turing Test
    - Codes
    - Very powerful

PROMPT

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

OUTPUT

They also were found to have perfectly coiffed hair, and wore what appeared to be Dior makeup.

“We were shocked to discover the unicorns,” said anthropologist Daniel St. Maurice. “They were like nothing we had ever seen before. We had heard legends of the unicorns, but never thought they actually existed.”

When the scientists first arrived in the valley, the unicorns were surprised and startled by the presence of humans, but were also excited. The unicorns welcomed the researchers and explained that they had been waiting for them for a very long time.

“The unicorns said that they had been waiting for us for a very long time,” said Dr. St. Maurice. “They said they had always known that humans would eventually discover them, but that they had also always known that humans would be too stupid to realize the unicorns had been waiting for them.”

The unicorns told the scientists that they had left the valley to explore the world, but that after years of traveling the world, they found humans to be incredibly stupid and decided to return to the valley to live in peace.

“We were shocked,” said Dr. St. Maurice. “We had no idea that the unicorns had been waiting for us. We also had no idea that unicorns were capable of speech. It was a truly magical moment.”

“We were shocked,” said Dr. St. Maurice. “We had no idea that the unicorns had been waiting for us. We also had no idea that unicorns were capable of speech. It was a truly magical moment.”

The unicorns said that they had long ago become tired of humans, but that they were excited to learn about the progress that had been made in the last 100 years.

“They were also disappointed to learn that in the past 100 years, we had not discovered the cure for cancer,” said Dr. St. Maurice. “They also seemed shocked that we were still fighting in wars.”

After a day of interacting with the unicorns, the scientists were eager to return to civilization and tell everyone about their amazing discovery. However, the unicorns had other plans.

“We wanted to return to civilization to tell everyone about the unicorns, but the unicorns refused to let us leave,” said Dr. St. Maurice. “They said they were very disappointed that we were so stupid and that we had wasted our lives on Earth.”

The unicorns then asked the scientists to help them kill all of the humans on Earth. The scientists agreed, saying that they were

Codex

- Watch video: https://openai.com/blog/openai-codex

```
#Python 2.7

#You are given a string. Split the string on a " " (space) delimiter
and join using a - hyphen.
```

And it spat out a Python function, which I pasted into my Aptana Python IDE. I made one update to Codex's output, changing an `input()` to a `raw_input()`.

```
1  def split_and_join(line):
2      line=line.split()
3      line="-".join(line)
4      return line
5
6  if __name__ == '__main__':
7      line = raw_input()
8      result = split_and_join(line)
9      print(result)
```

source

CLIP

FOOD101

guacamole (90.1%) Ranked 1 out of 101 labels

✓ a photo of **guacamole**, a type of food.

× a photo of **ceviche**, a type of food.

× a photo of **edamame**, a type of food.

× a photo of **tuna tartare**, a type of food.

× a photo of **hummus**, a type of food.

SUN397

television studio (90.2%) Ranked 1 out of 397

✓ a photo of a **television studio**.

× a photo of a **podium indoor**.

× a photo of a **conference room**.

× a photo of a **lecture room**.

× a photo of a **control room**.

YOUTUBE-BB

airplane, person (89.0%) Ranked 1 out of 23

✓ a photo of a **airplane**.

× a photo of a **bird**.

× a photo of a **bear**.

× a photo of a **giraffe**.

EUROSAT

annual crop land (12.9%) Ranked 4 out of 10

× a centered satellite photo of **permanent crop land**.

× a centered satellite photo of **pasture land**.

× a centered satellite photo of **highway or road**.

✓ a centered satellite photo of **annual crop land**.

Robotics

Etc. etc. etc.

- Go
- Dota
- Starcraft
- Etc..