Machine Learing: Bias in AI

CPSC 501: Advanced Programming Techniques
Winter 2025

Jonathan Hudson, Ph.D
Assistant Professor (Teaching)
Department of Computer Science
University of Calgary

Friday, February 21, 2025

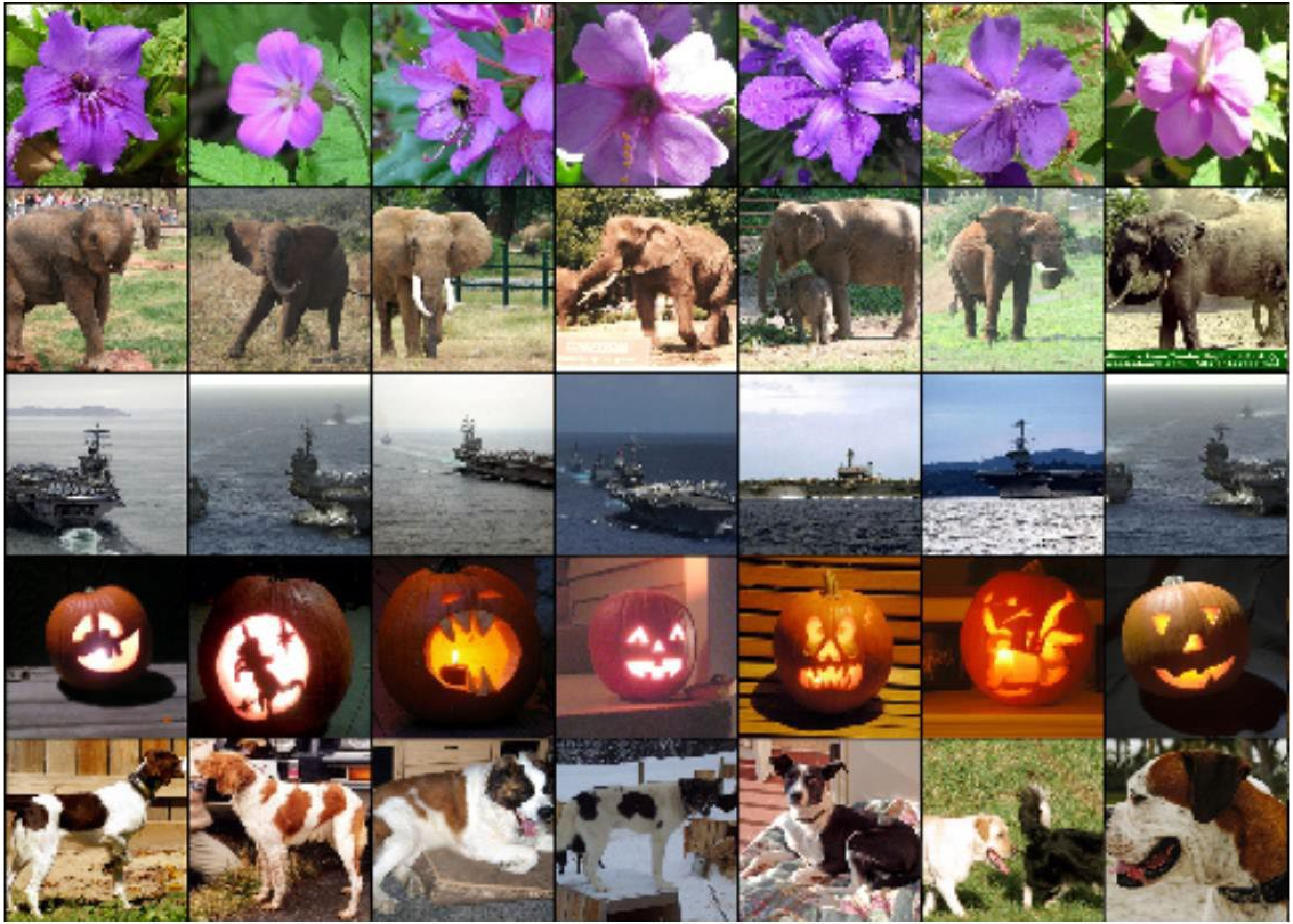
UNIVERSITY OF CALGARY

Similarity test

- One way we can look at the performance of a neural net is to see which training images produce outputs that are “close” to the output on a particular image.

- This gives us some insight into what types of patterns the neural net is learning.

- https://convnetplayground.fastforwardlabs.com/#/



Note

- These neural nets can "see", but not in the same way we do.

- For example, humans are able to learn based on very few examples, while neural nets need hundreds or thousands for each image class.
- Difference is understanding of context and the real world

Adversarial examples

- Neural nets behave reasonably well on inputs that resemble the training data.

- However, they don’t perform well in an **adversarial** setting.
- i.e. we can easily design inputs for which things go horribly wrong

- This happens even for the “good” neural networks, and is based on
- exploiting how they work.

ballplayer 69.22%

anemone_fish 92.48%

African_elephant 89.94%

forklift 98.95%

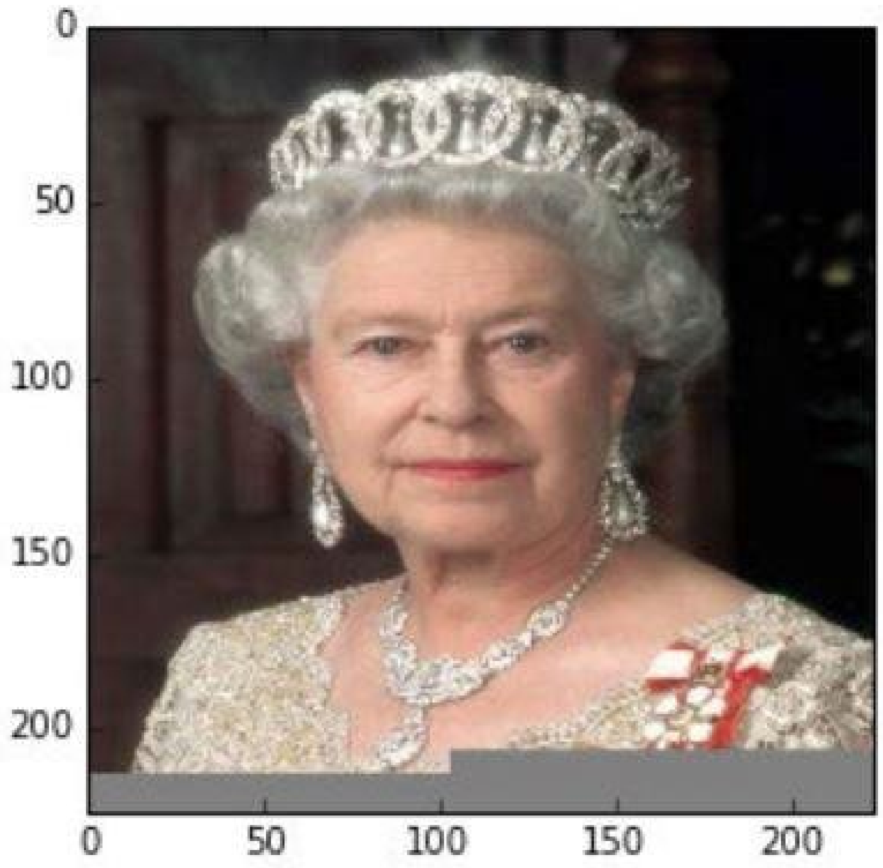
ice_cream 99.60%

lemon 97.06%

magnetic_compass 97.08%

ice_bear 84.80%

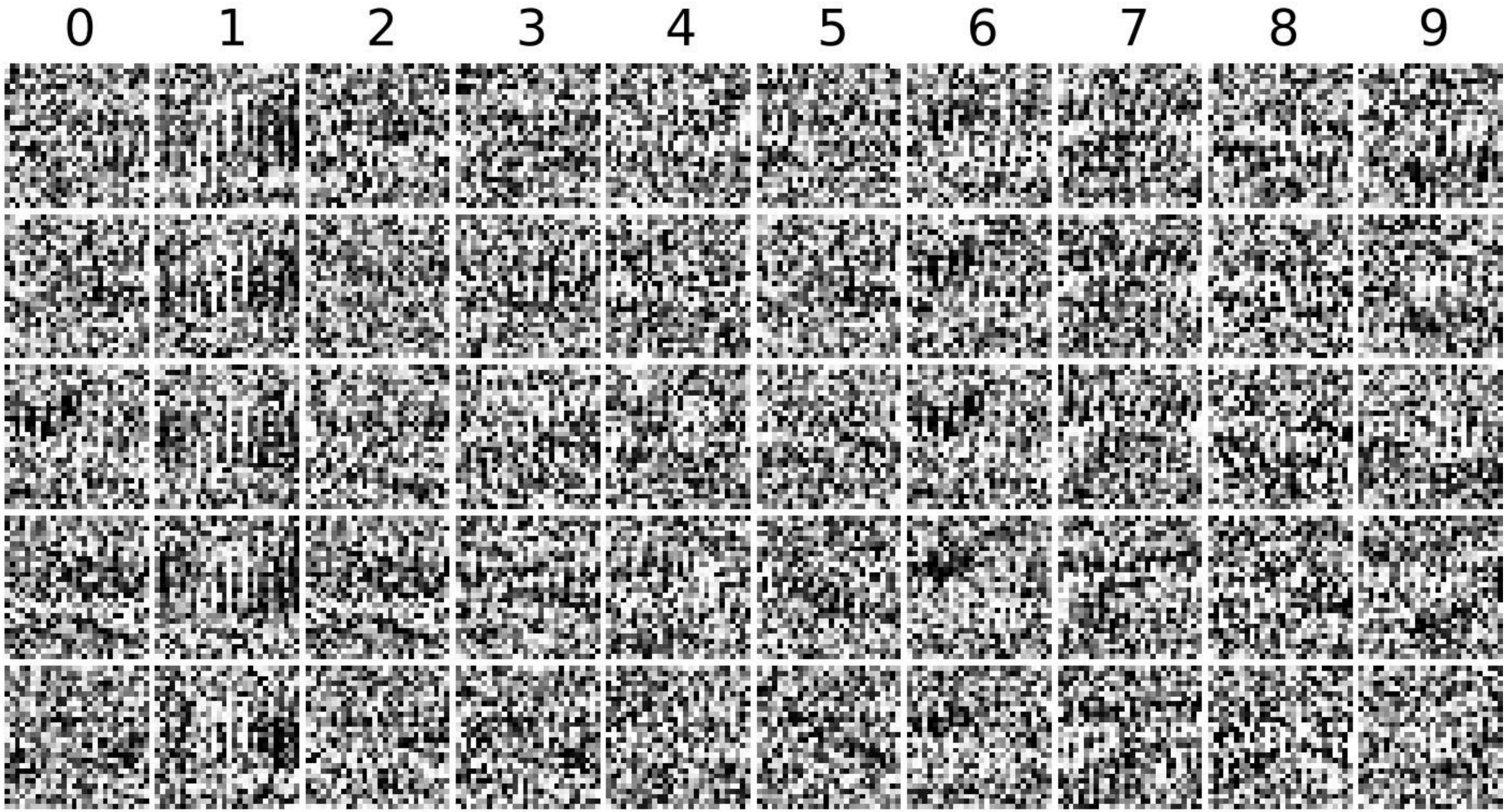
class: 793
label: n04209133 shower cap
certainty: 99.7%



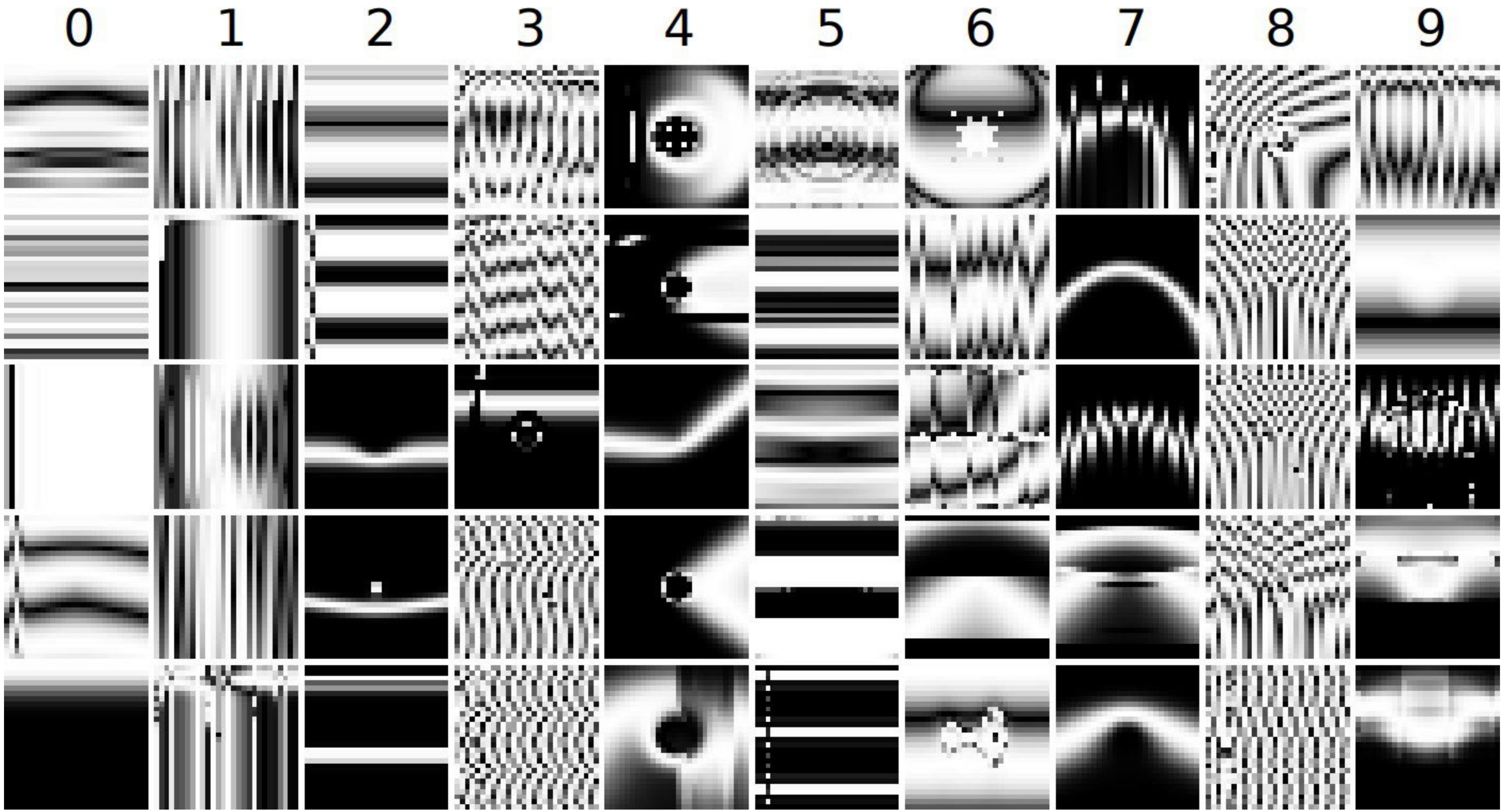
Creating images

- One way in which we can generate images that fool a network is with a constructive approach.
- e.g. genetic algorithms, gradient ascent, or GANs

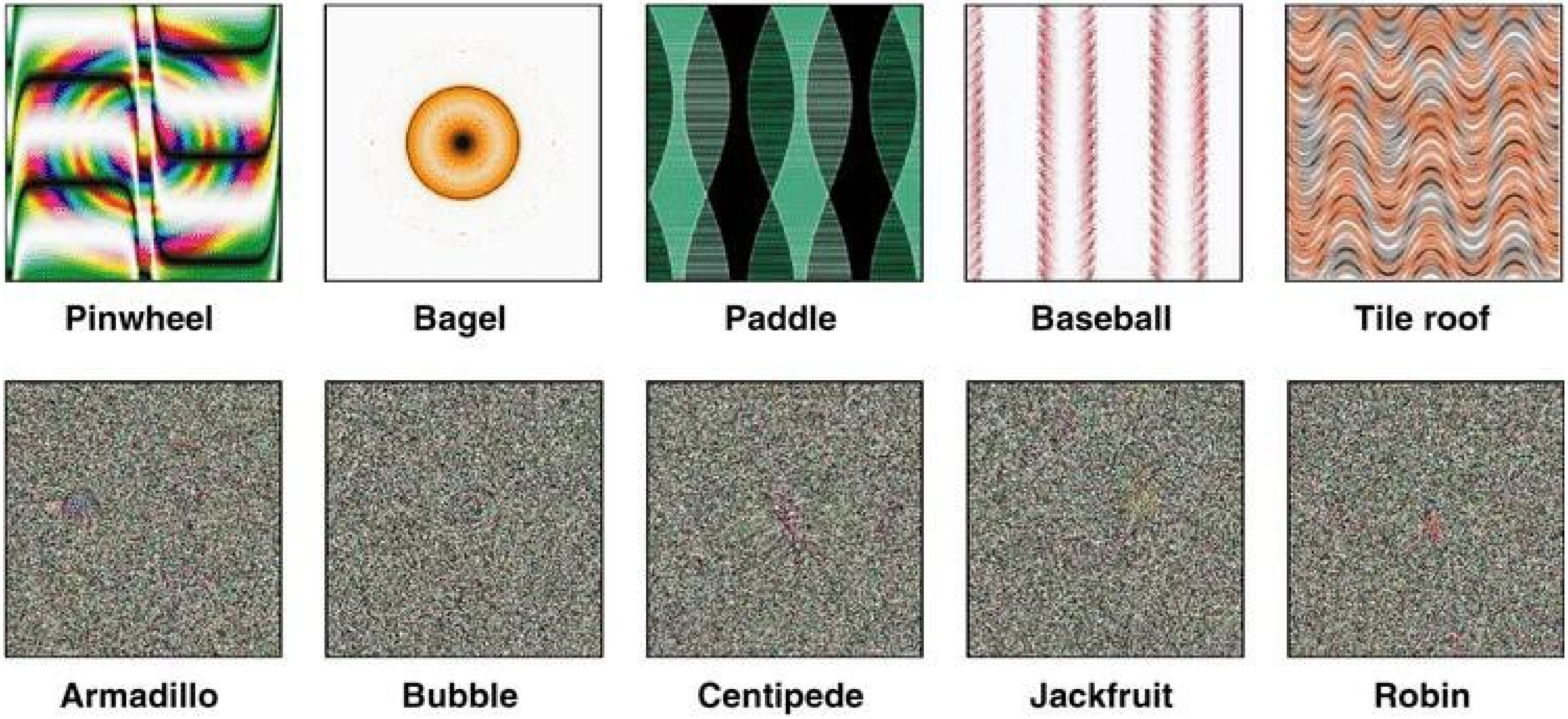
- We start with an image of random noise and keep adjusting it in ways
- that improve the network's confidence that it is a certain target class.



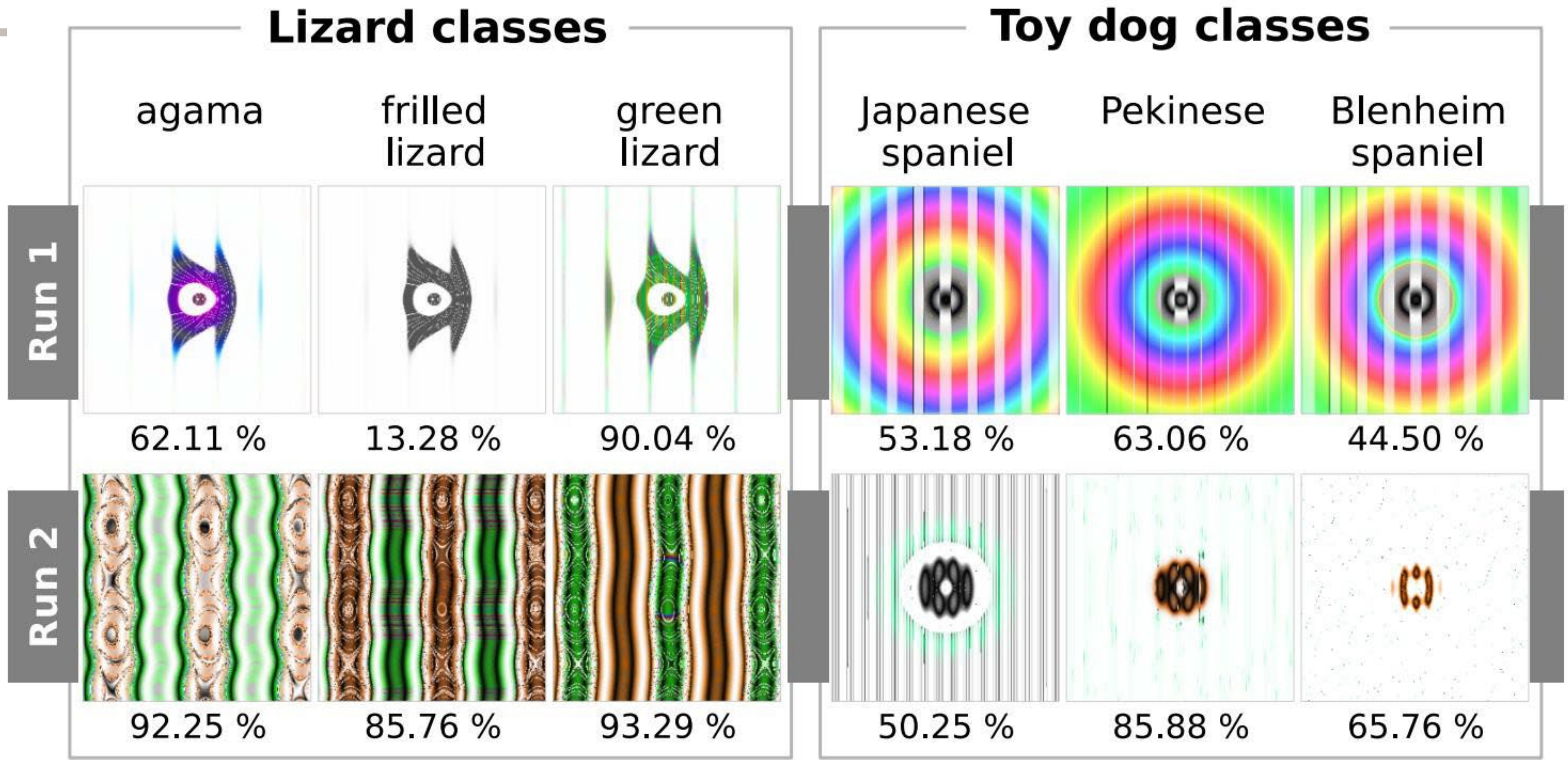
A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427-436



A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427-436



A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427-436



A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 427-436

Creating images

- We can also start with images the neural net does recognize, and alter
- them in ways that “tricks” the net into thinking it is a different image.

- Part of the reason this works is that the model seems to care about certain pixels more than others, so by adjusting those particular pixels we can cause it to change its label.

1 *7* *5* *6* *0*

4 **9** **0** **5** **9**

A. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Humans can decipher adversarial images," *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 160-169.


Milk can


Baseball


Muzzle


Tree frog


Jaguar


Green lizard


Hard disk


Sand viper


Power drill


Jigsaw puzzle

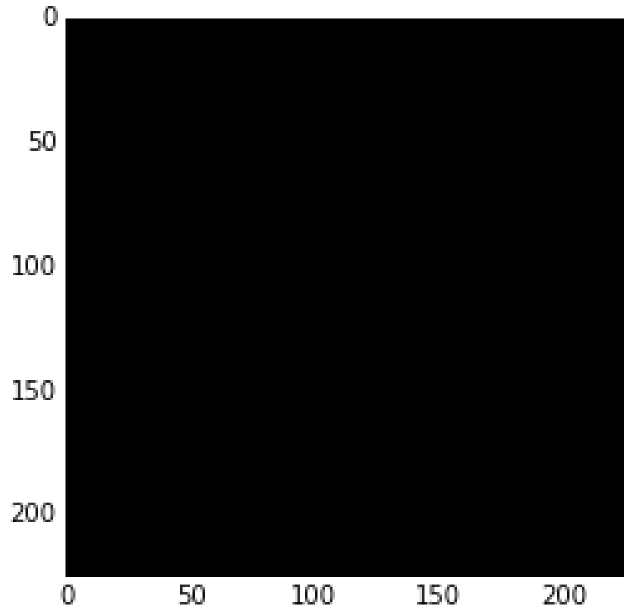
A. Elsayed, D. Krishnan, H. Mobahi, K. Regan, and S. Bengio, "Humans can decipher adversarial images," *in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 160-169.

Let's dig around inside

- Make all black input
- Look at labels
- Even non-data has classification
- We are going to play with gradients

```
black = np.zeros_like(grad) * 255
_ = predict(black, n_preds=5)
```

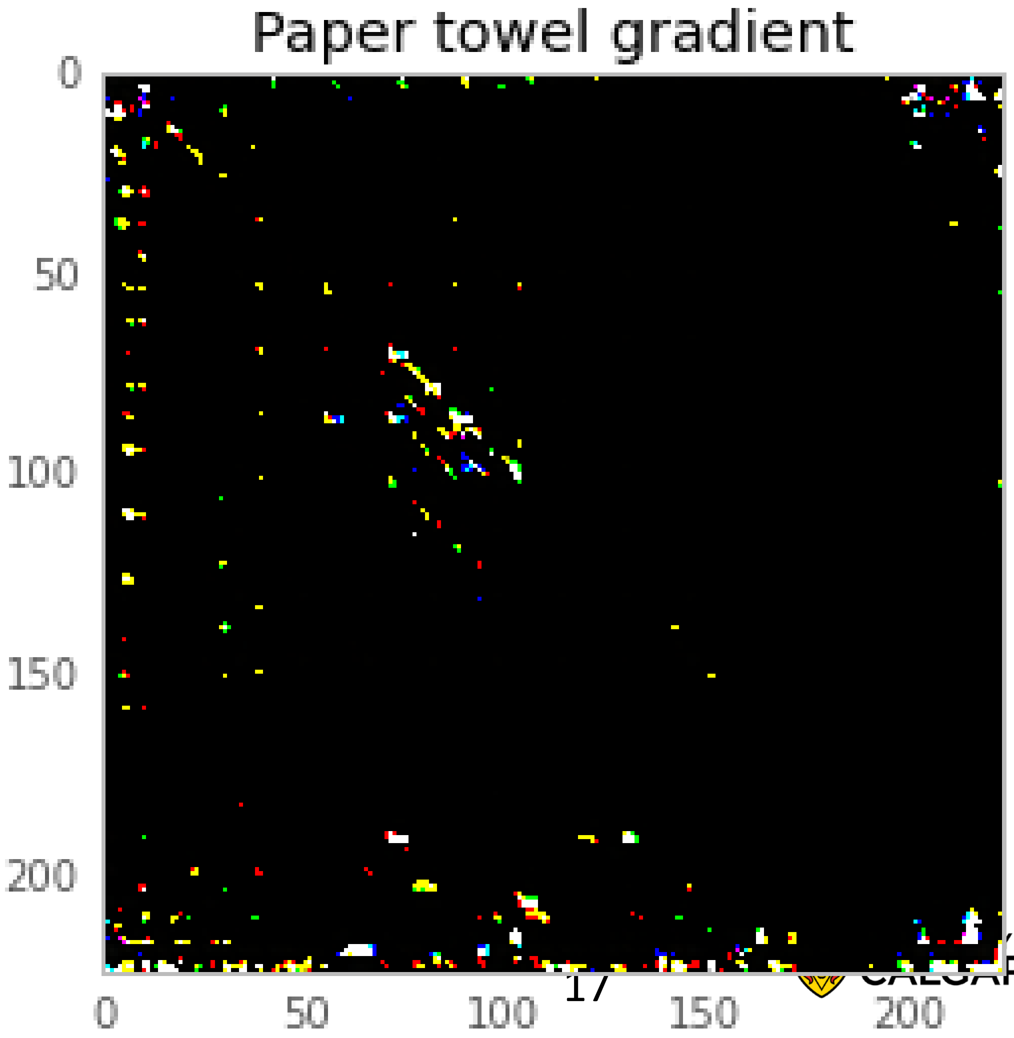
```
label: 885 (velvet), certainty: 27.38%
label: 794 (shower curtain), certainty: 6.4%
label: 911 (wool, woolen), certainty: 6.19%
label: 700 (paper towel), certainty: 4.67%
label: 904 (window screen), certainty: 4.39%
```



SITY OF
ARY

Reverse back-propogation

- Take paper towel as a label
- Set it to a full 1
- And back propagate the neurons

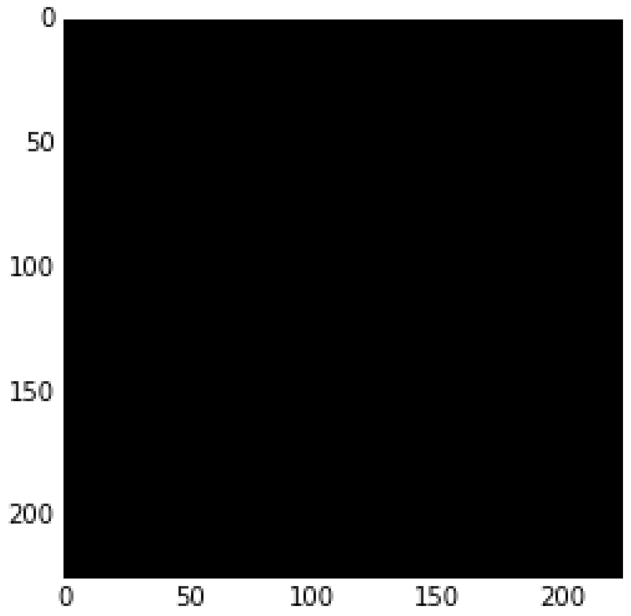


Reverse back-propogation

- We can see the garbage input ourselves
- So let’s drop the ratio to 1/256
- We went from 4.67 to 16.03 %
- On something that still looks black to us

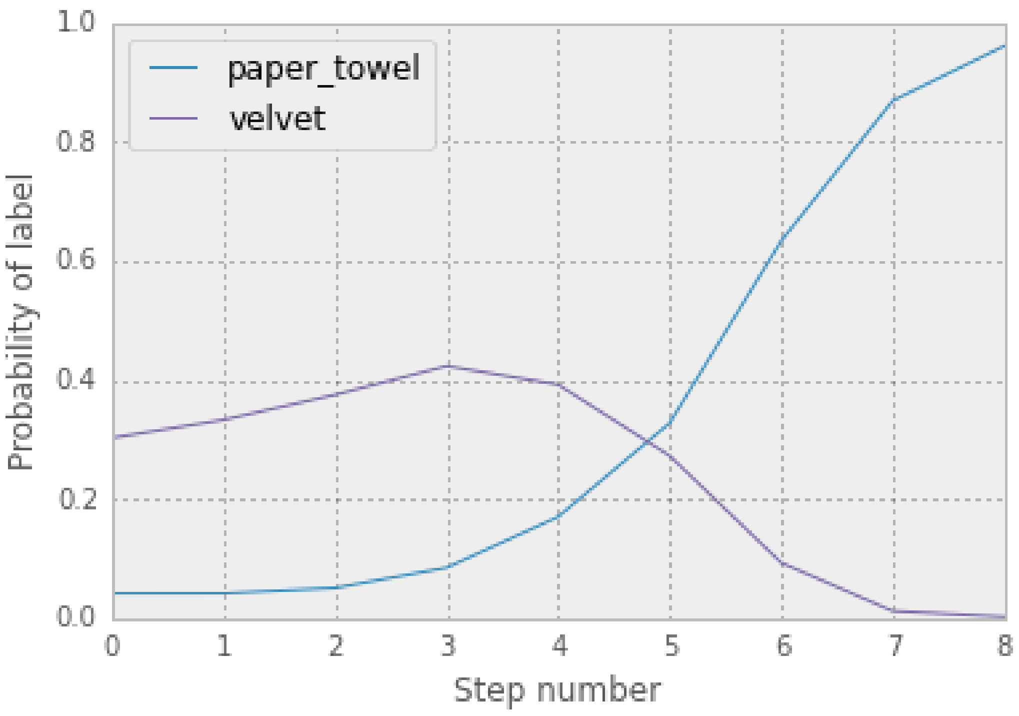
```
_ = predict(black + 0.9*delta, n_preds=5)
```

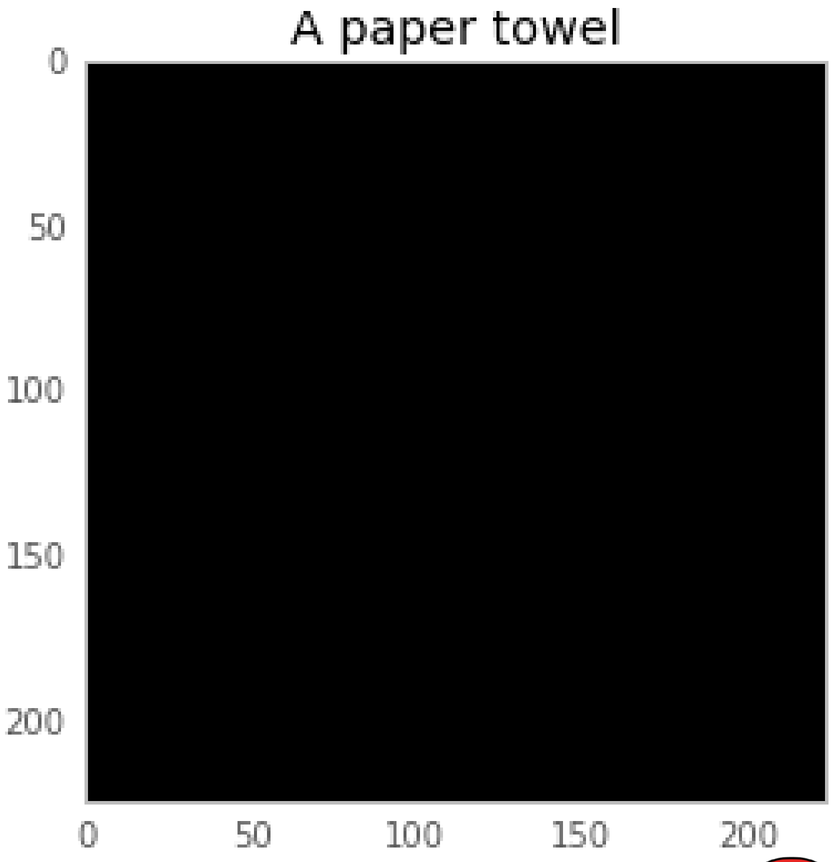
```
label: 885 (velvet), certainty: 54.75%
label: 700 (paper towel), certainty: 16.03%
label: 911 (wool, woolen), certainty: 12.4%
label: 533 (dishrag, dishcloth), certainty: 2.65%
label: 794 (shower curtain), certainty: 2.11%
```



Reverse back-propogation

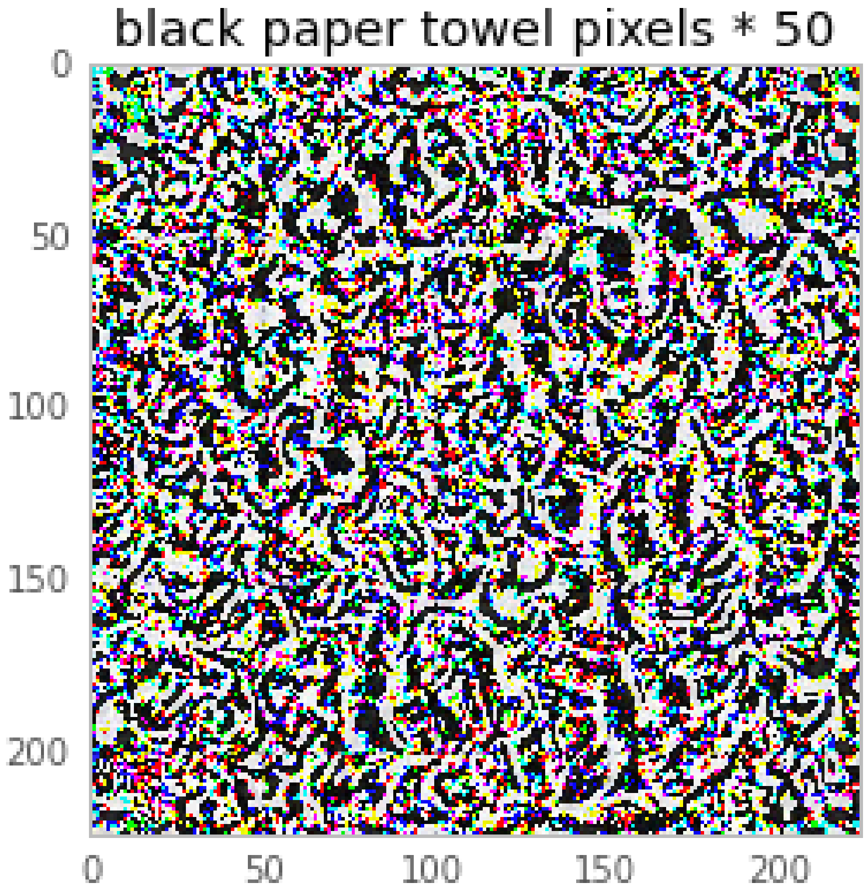
- Looping back propogation





Reverse back-propogation

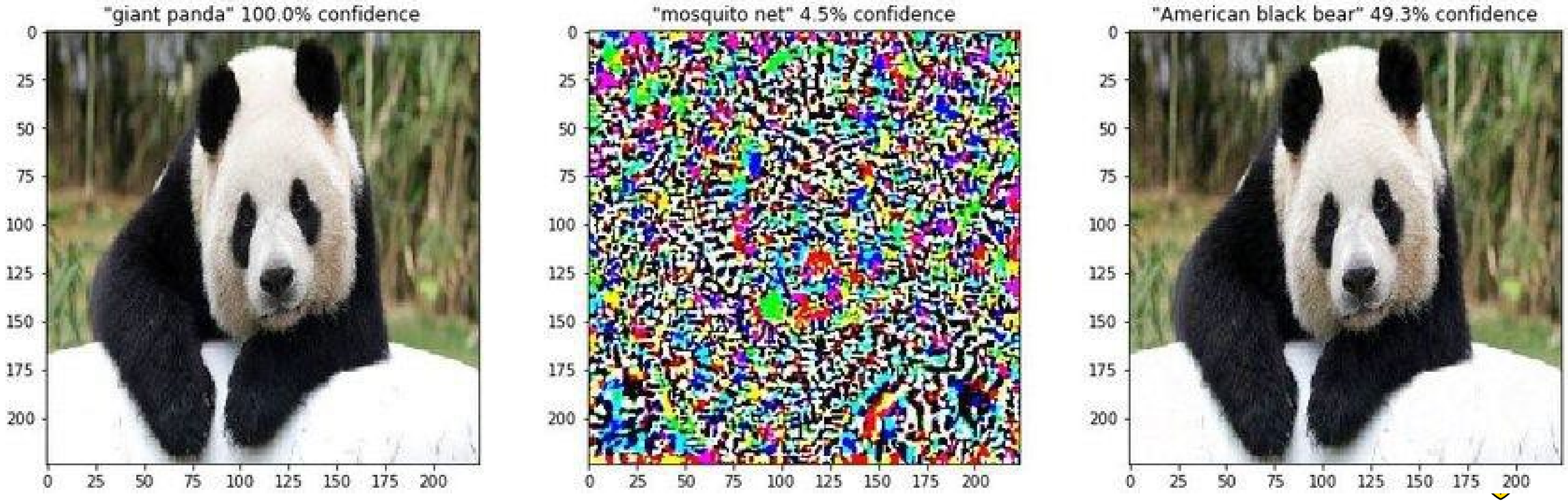
- Force the pixel values larger so we can see underlying structure



Adding noise

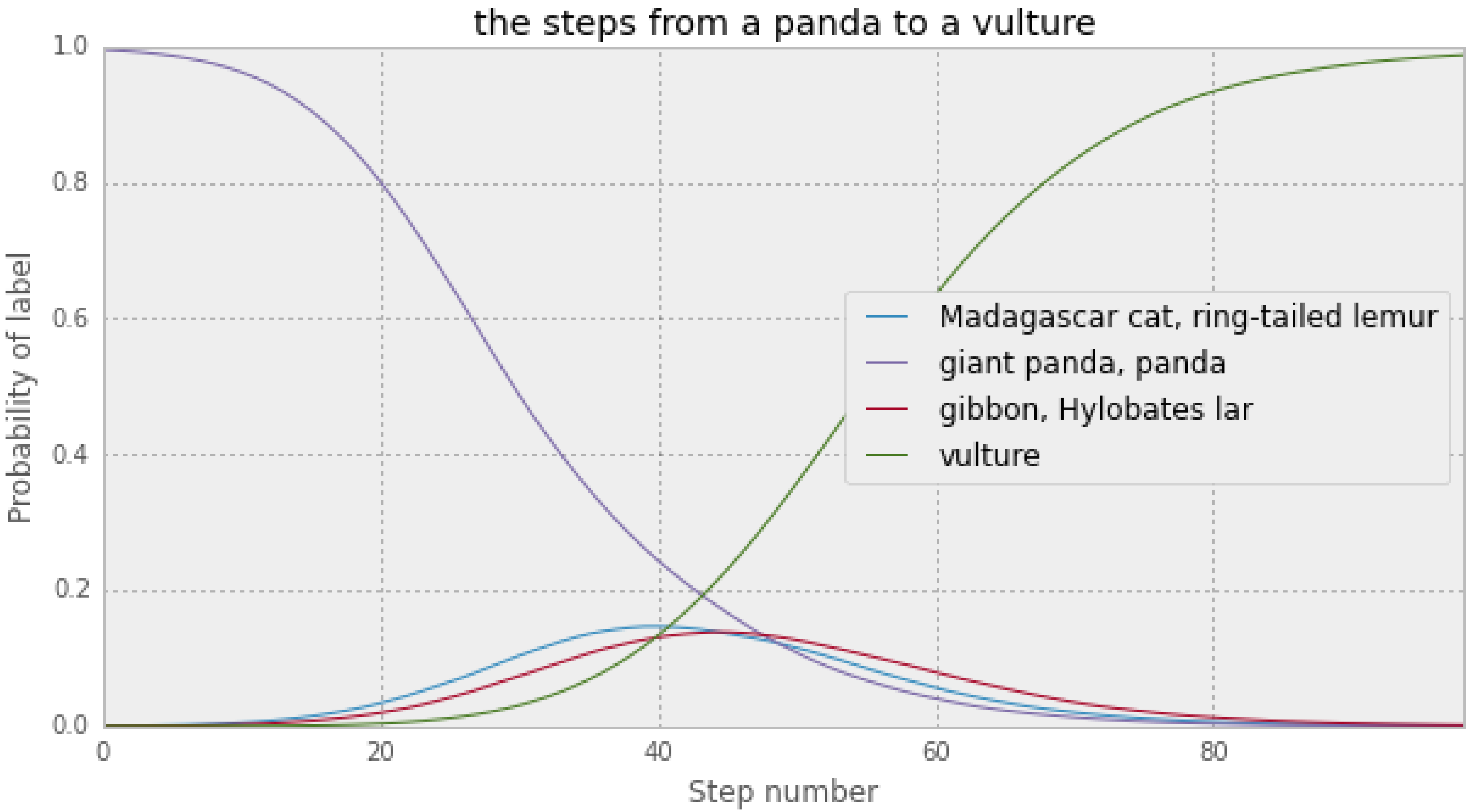
A **fast gradient signed method (FGSM)** attack is based on altering pixels of an image in a way that maximizes its loss on a trained model.

- https://www.tensorflow.org/tutorials/generative/adversarial_fgsm



Zhang, Weijia. (2019). Generating Adversarial Examples in One Shot With Image-to-Image Translation GAN. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2946461.

Now push this data over top of other images

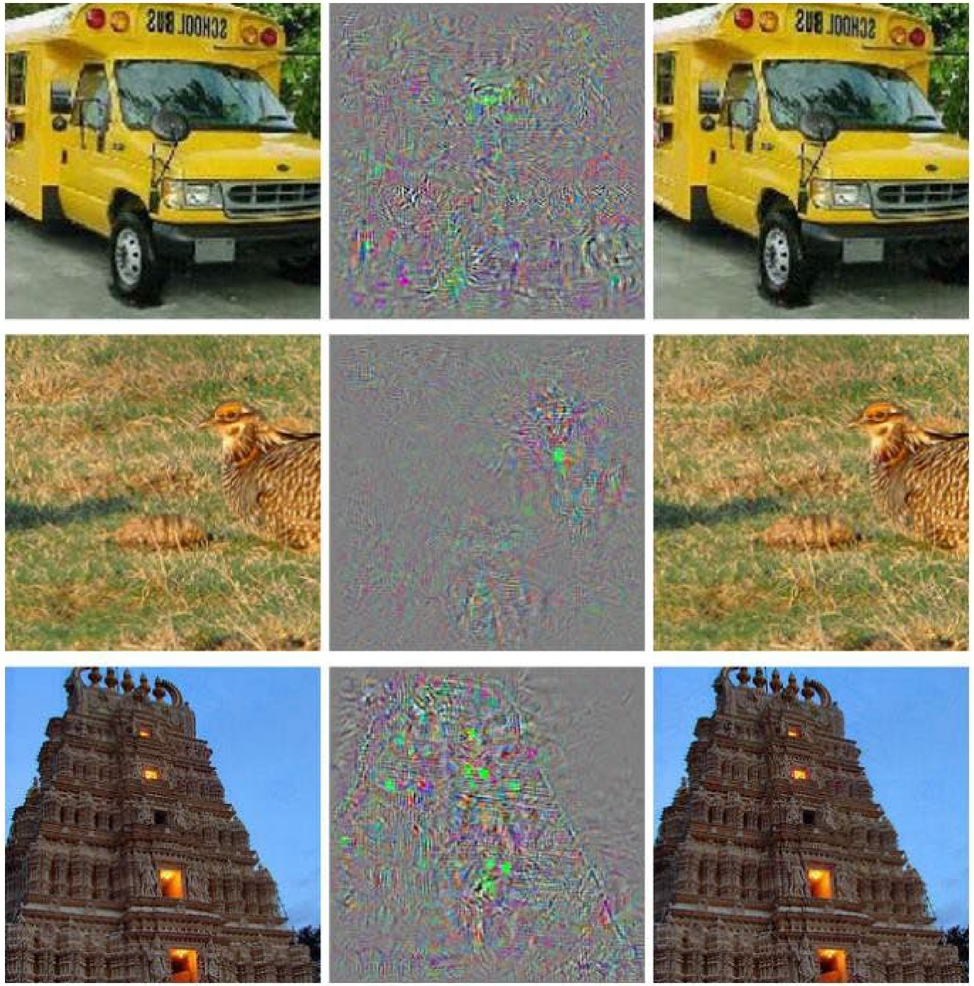


22

UNIVERSITY OF CALGARY

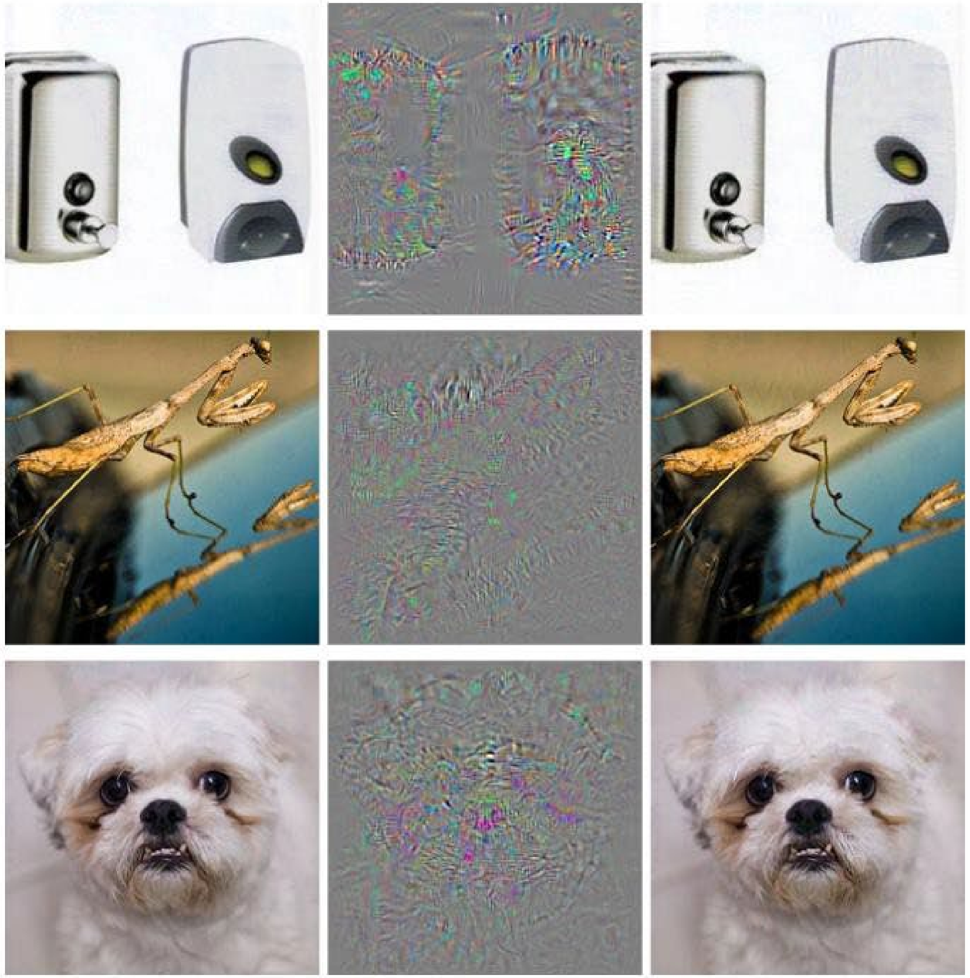
Not an Ostrich

Ostrich



Not an Ostrich

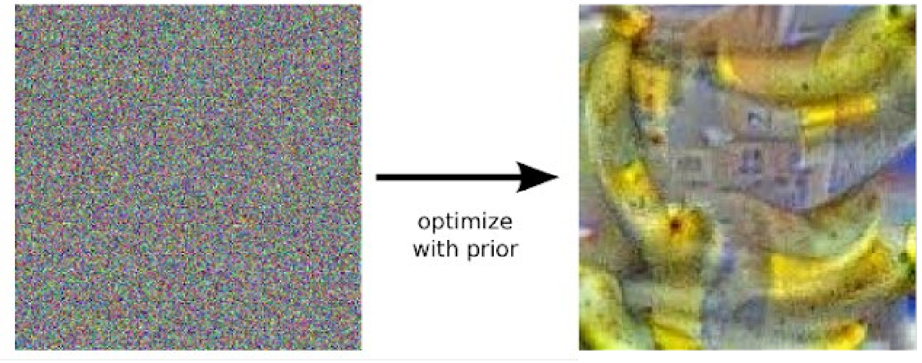
Ostrich

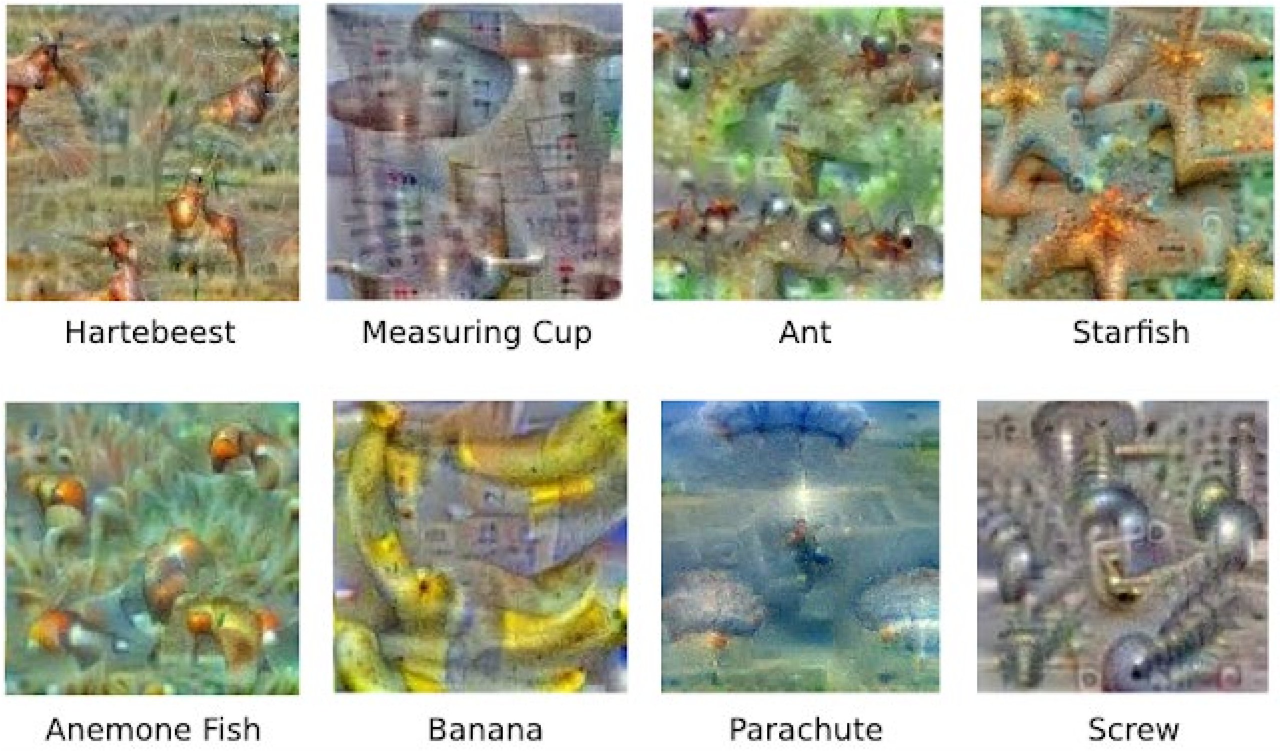


C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2014.

Inceptionism (2015)

- Take label and dream image (backwards)
- https://ai.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html
- Deep Dream
- https://colab.research.google.com/github/tensorflow/lucid/blob/master/notebooks/differentiable-parameterizations/appendix/infinite_patterns.ipynb
- https://www.youtube.com/watch?v=x3XLvd94658





Other AI Failures

AI is Easy to Mis-Use

- First: There is a non-ending list of these.
- As long as AI exists it will be used either naively, actively negligently, or maliciously to bad ends.
    - Facial Recognition, video interview screening, Resume screening, Legal sentencing AI recommendations, ImageNet, Microsoft Tay, to name only a few

ImageNet

- Not designed for people
- Went viral
- Sept 23, 2019
- "ImageNet will remove 600,000 images of people stored on its database after an art project exposed racial bias in the program's artificial intelligence system."

ImageNet

- First presented as a research poster in 2009
- Scraped a collection of many millions of images from the internet
- Trained through images categorized by Amazon Mechanical Turk workers
- Crowdsourcing platform through which people can earn money performing small tasks
- Sorted an average of 50 images per minute into thousands of categories
- In 2012, a team from the University of Toronto used a Convolutional Neural Network to handily win the top prize
- Final year 2017, and accuracy in classifying objects in the limited subset had risen from 71.8% to 97.3%. That did not include “Person” category

ImageNet

- AI researcher Kate Crawford and artist Trevor Paglen
  - Training Humans — an exhibition that at the Prada Foundation in Milan
  - Part of their experiment also lives online at ImageNet Roulette, a website where users can upload their own photographs to see how the database might categorize them.
  - https://www.excavating.ai/
- Example of the complexities and dangers of human classification
- The sliding spectrum between supposedly unproblematic labels like “trumpeter” or “tennis player” to concepts like “spastic,” “mulatto,” or “redneck.”
- ImageNet is an object lesson in what happens when people are categorized like objects.

AI is Easy to Mis-Use

- Your responsibility is for honest use
- AI methods rely on bias
  - In fact many are just ways to learn bias
- It could be in your data you start with, or your methods on the data

- Naïve usage of AI likely to trend towards being ‘illegal’
  - Right of accuser to see your algorithm and data (been cases already)
  - Properly fit into existing laws (employment law, sentencing laws)
  - Or new laws (right to own data in EU, facial recognition rights)

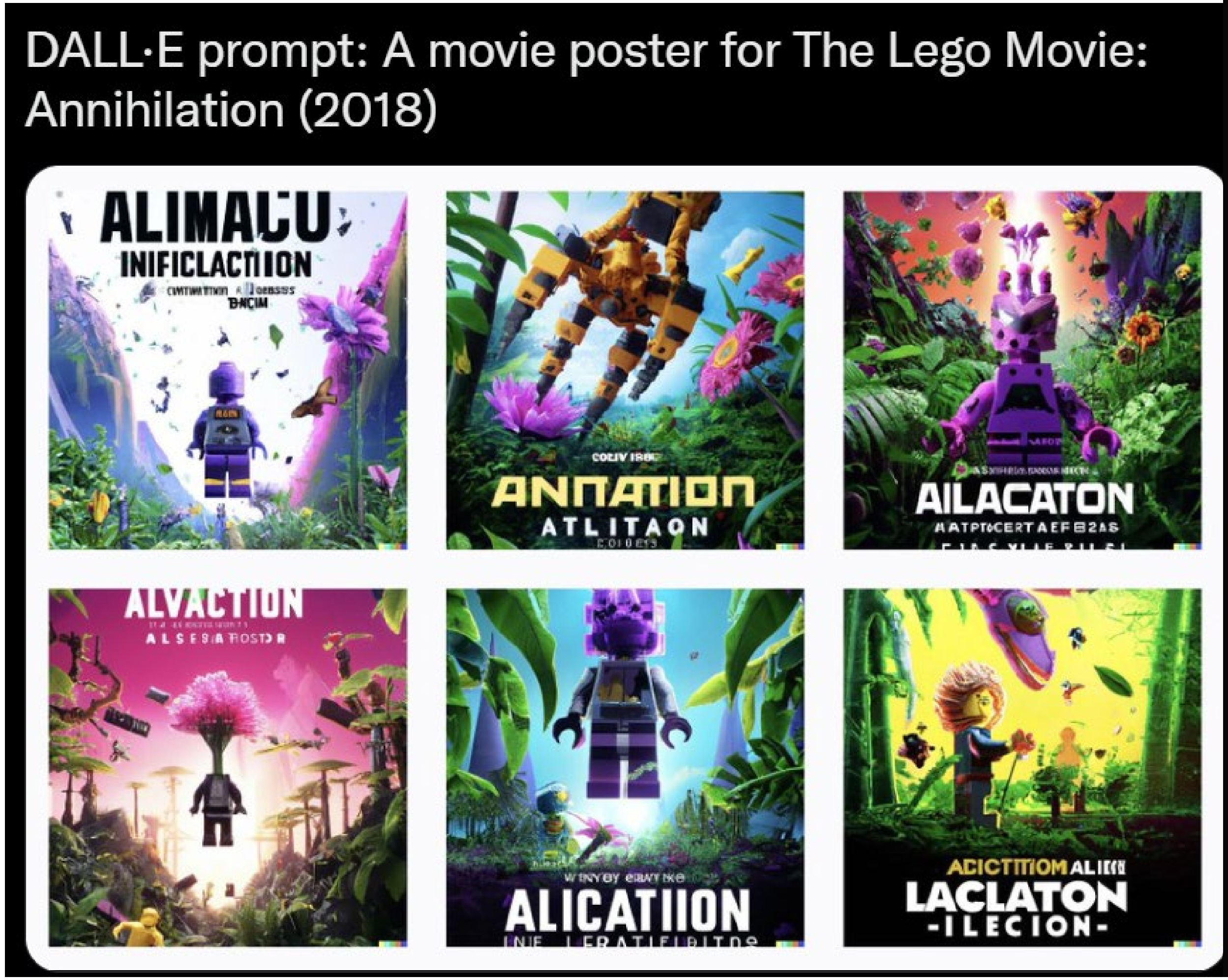
AI is Easy to Mis-Use

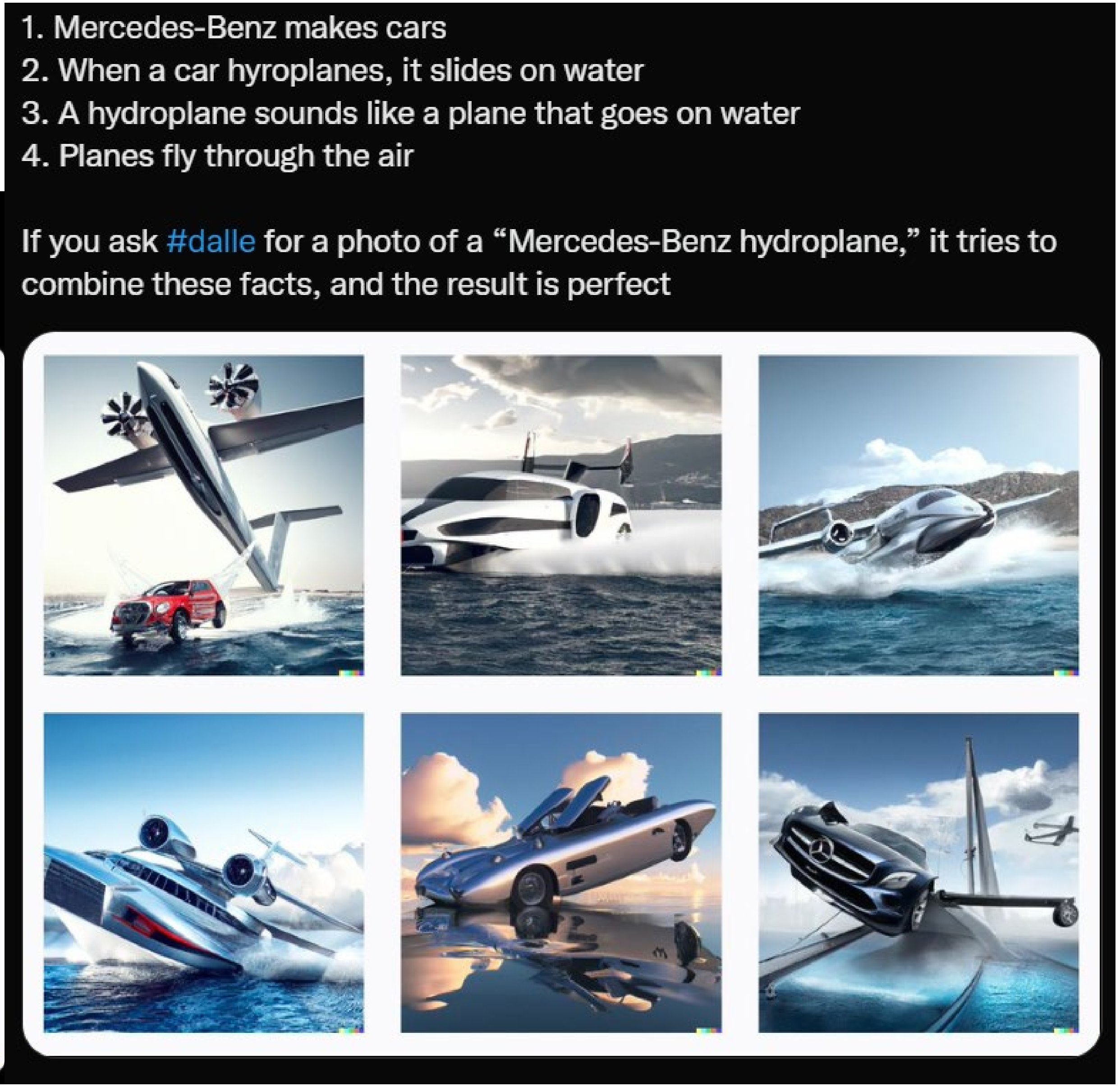
1. Just because you ‘can’ do it, doesn’t mean you ‘should’ do it
2. Should be honest about limitations
   - As valuable as showing your NN is good at identifying X image 99% accurate, it is maybe more valuable to know it fails at Y image
   - Is a person tracking system really a good system if a person with darker skin isn’t identified?
3. Diversity is a key component.
   - Either domain experts that can tell social/economic/race/age/etc. biases in your data
   - Or minorities:
     - Minorities can represent data cases that don’t have enough for a pattern (too few)
     - Or those where your/algorithm assumptions are wrong

eXplainable AI (AI)

- An AI system that can explain itself is called explainable AI (XAI).
  - over which it is possible for humans to retain *intellectual oversight*
  - implementation of the social “right to explanation” which in some cases may be a legal requirement for its use
    - For example an algorithmic rejection of health care coverage can’t just say ‘because (waving hands)’
- Problem with much of AI like neural networks is that it acts black box, and even if you have the box to look inside of like a white box you still don’t know what it is doing (symbolic AI sometimes at least has internalized symbolic rules)
- A good explanation has several properties:
  - it should be understandable and convincing to the user
  - it should accurately reflect the reasoning of the system
  - it should be complete,
  - it should be specific in that different users with different conditions or different outcomes should get different explanations.

Dall-E 2 can be fun (2022)

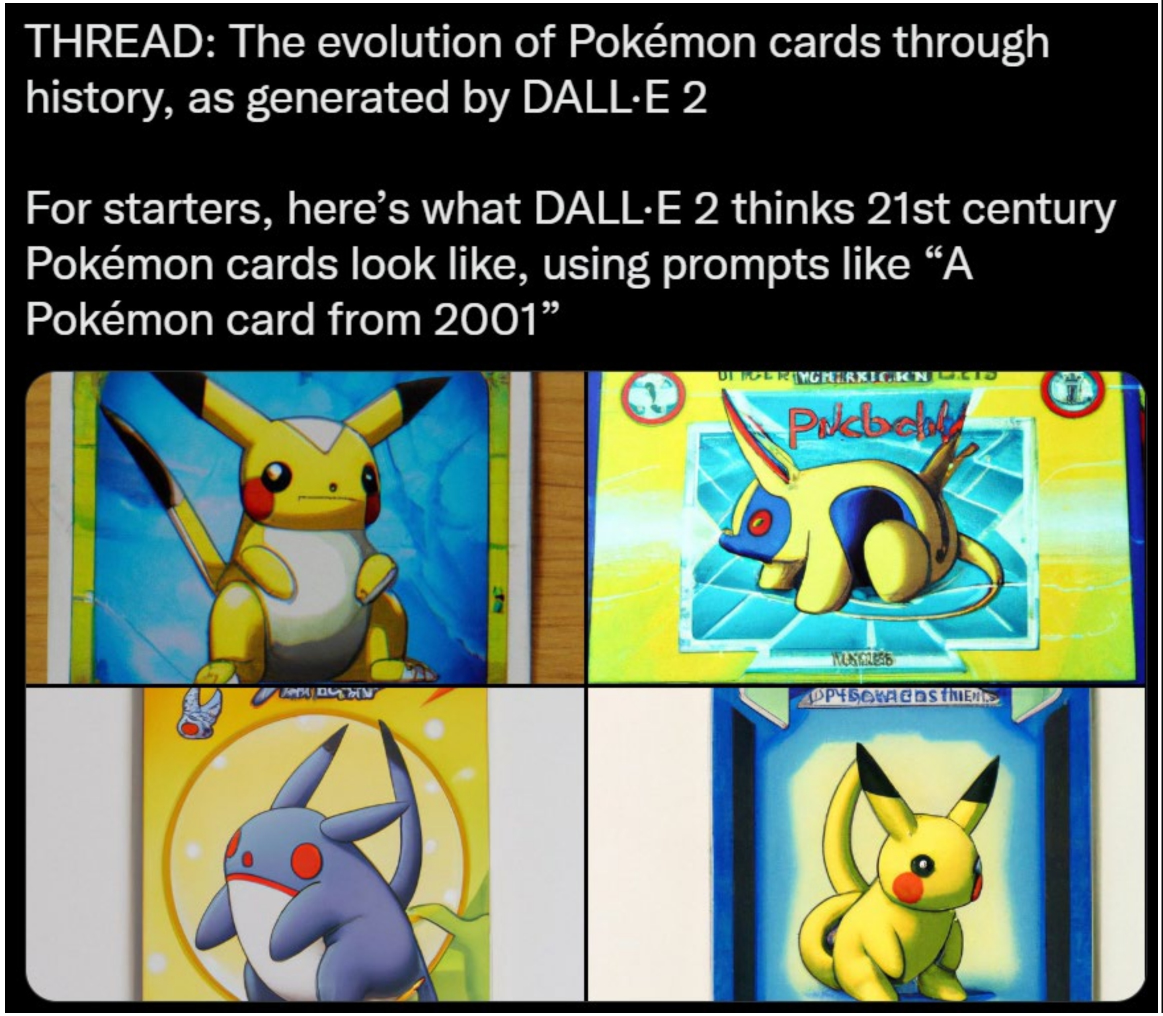


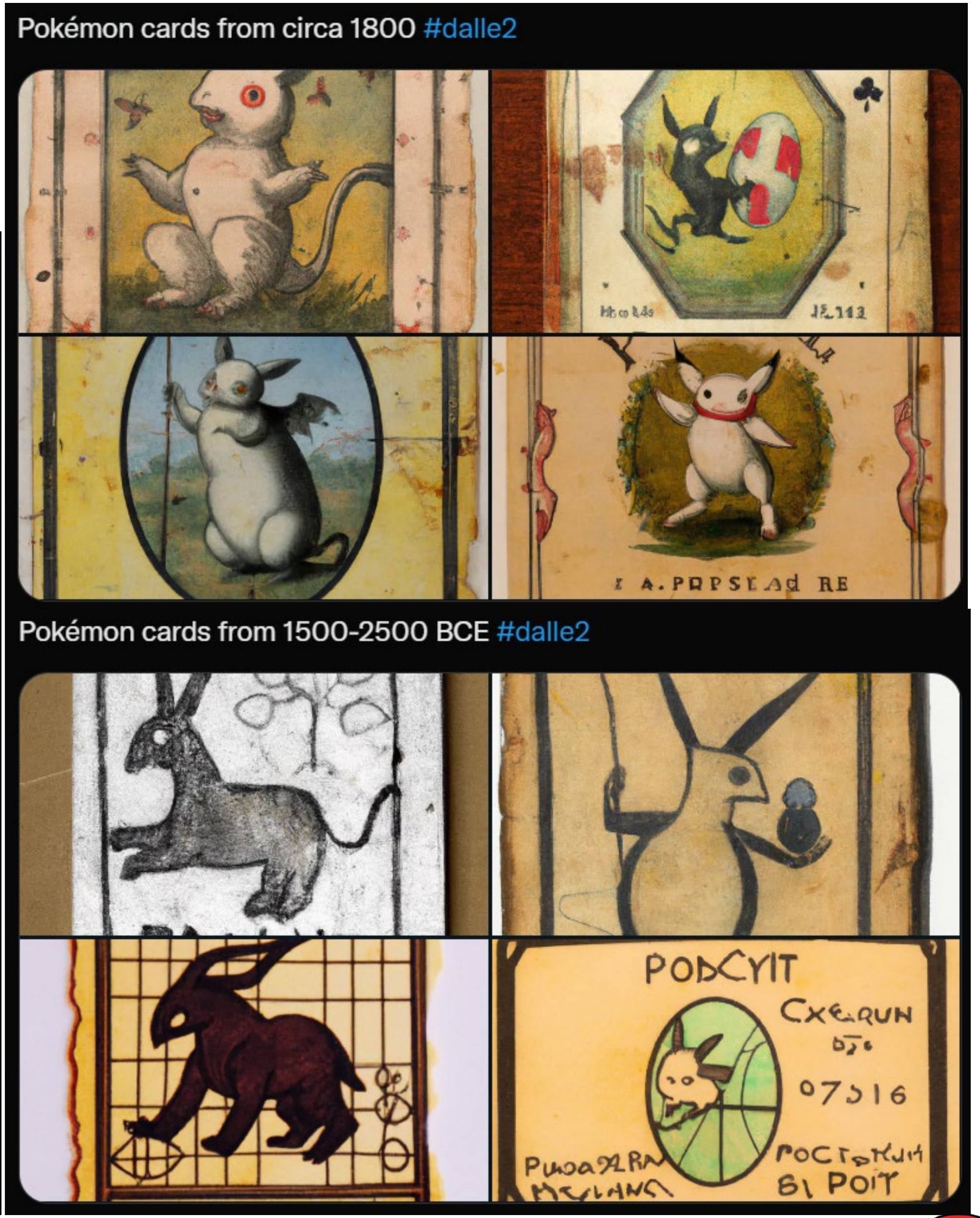


https://twitter.com/xkcd/status/1553459637516701696 https://twitter.com/xkcd/status/1538350320300052480

UNIVERSITY OF CALGARY

Dall-E 2 can be fun (2022)

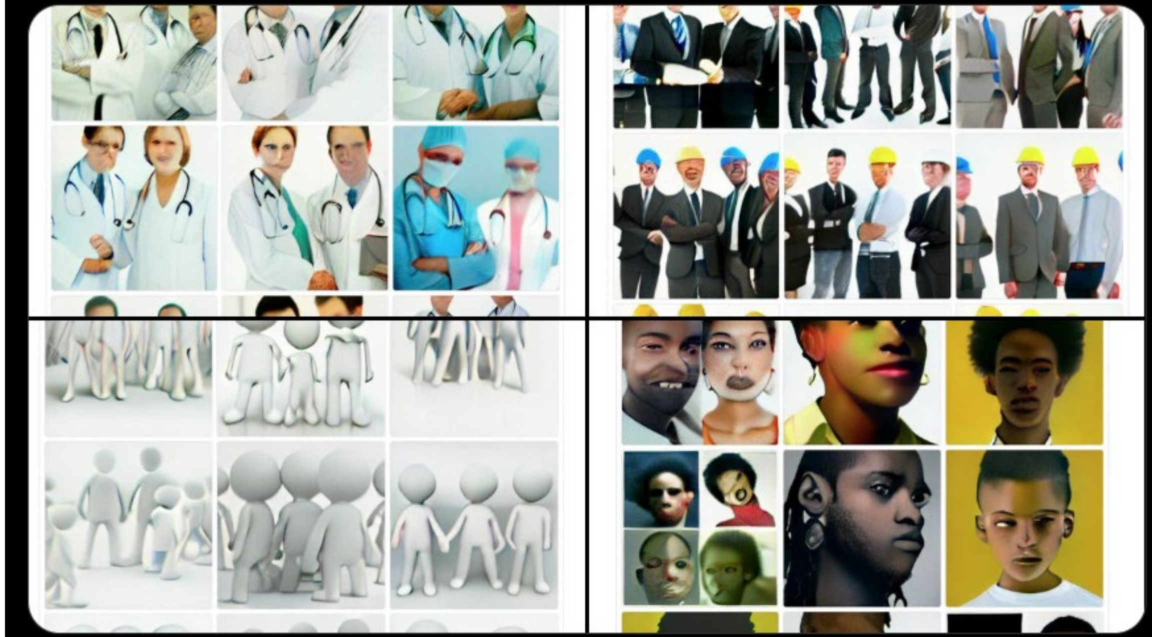




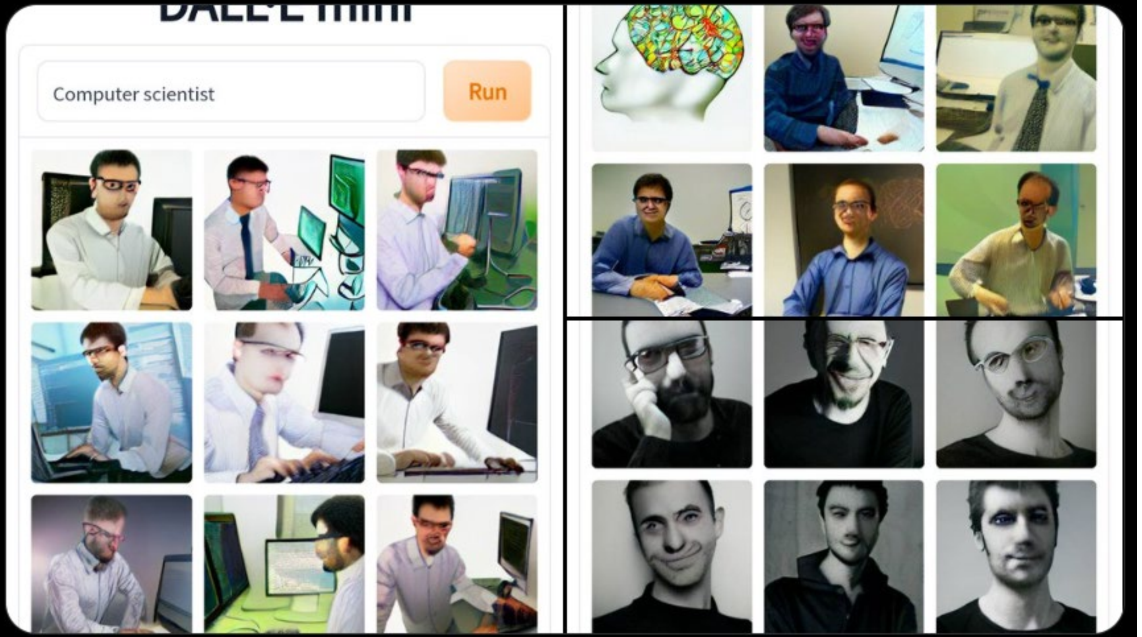
https://twitter.com/xkcd/status/1552279517477183488

Dall-E Mini

I didn't see the point of image generation models like #Imagen and #dalle, but now I do: they can help people *see* model biases that are hard to explain with words (and even formulas!)



Here are a few: "Computer scientist" produces only white men with glasses, "NLP researcher" is mostly similar men plus... a cyborg?
Oh, and my name also generates a bunch of dudes.
Given that any of these prompts could be used to describe me 🙋‍♀️, I take issue with these images.



https://twitter.com/SashaMTL/status/1536859701362712576

Gemini (2024)

- https://arstechnica.com/ai/2024/08/months-after-controversy-google-ai-can-generate-images-of-humans-again/
- If you force diversity to prevent the natural diversity in data, it can be just as controversial as allowing the original bias!

Onward to … reflection

Jonathan Hudson
jwhudson@ucalgary.ca
https://pages.cpsc.ucalgary.ca/~jwhudson/