

Data Science: Machine Learning: Dangers in AI

CPSC 501: Advanced Programming Techniques
Fall 2022

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

Monday, October 24, 2022



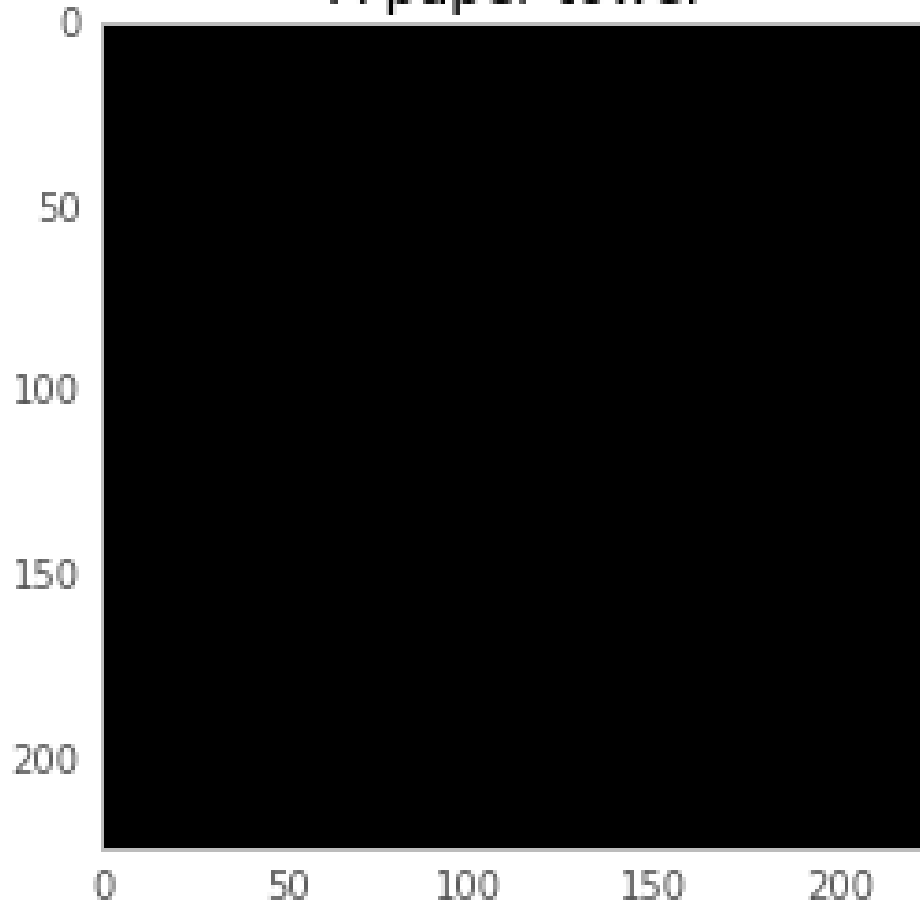
Neural Network Specific Adversarial Attacks

Adversarial

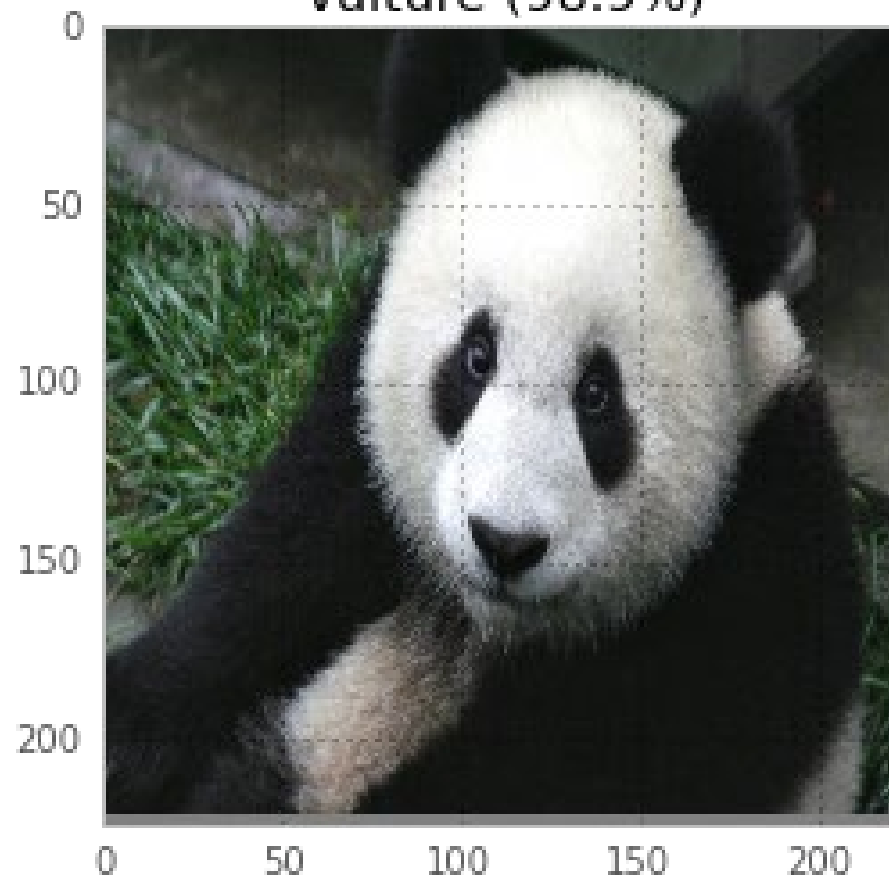
- We are usually nice to neural networks
- We feed it data like it has seen before and get back positive results
- Instead what if we are malicious and exploit how they work
- Main point (neural networks are basically very complicated functions which we can back solve and exploit)

What!

A paper towel



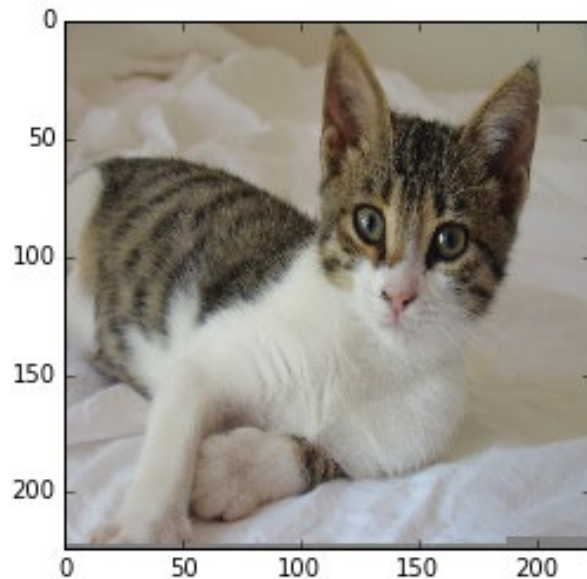
Vulture (98.9%)



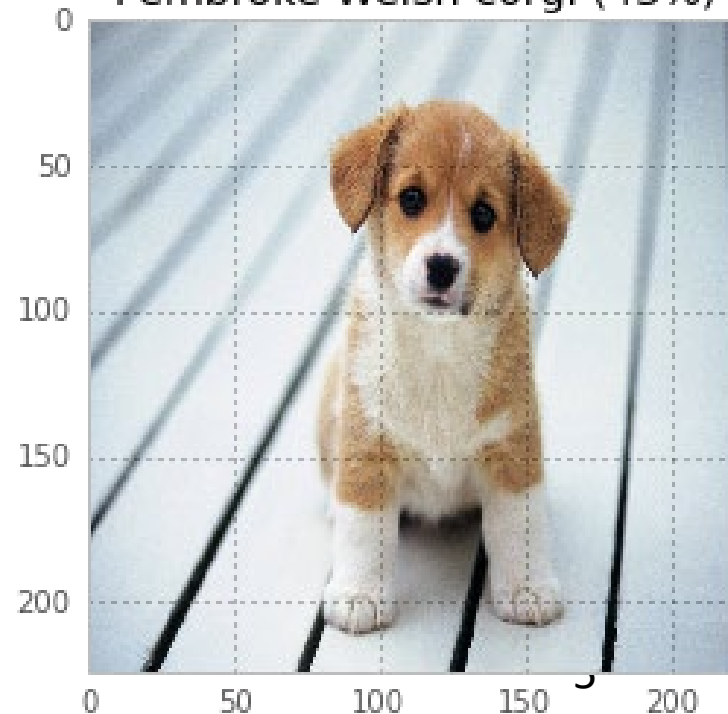
GoogLeNet

- These examples are GoogLeNet from the ImageNet competition
- So a large CNN, that should be (slightly) more robust against exploitation

class: 285
label: n02124075 Egyptian cat
certainty: 34.57%



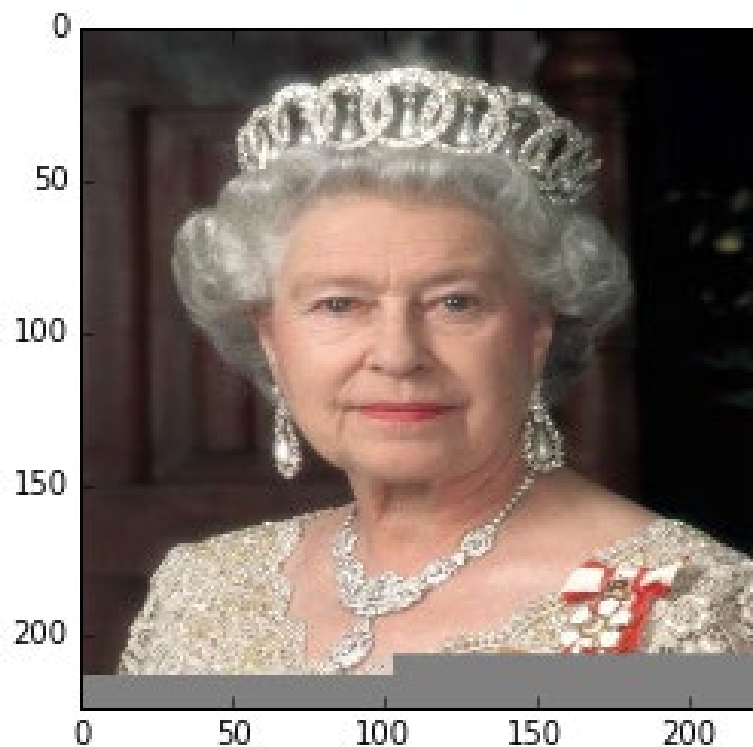
Pembroke Welsh corgi (43%)



GoogLeNet

- Even good CNNs can suck without our help

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

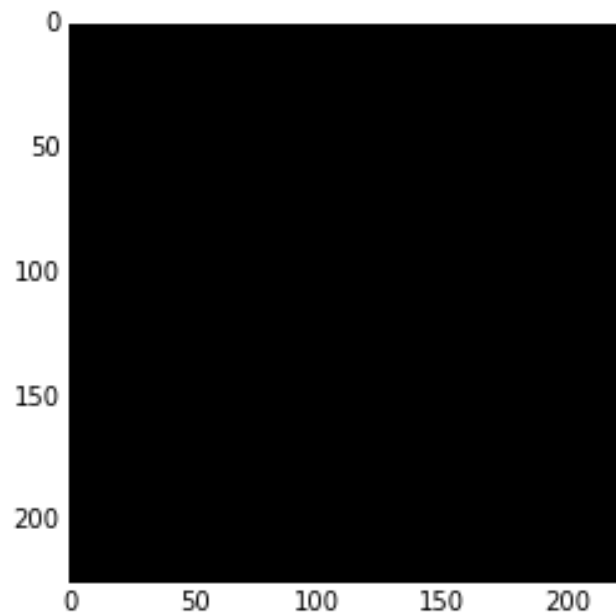


Let's dig around inside

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

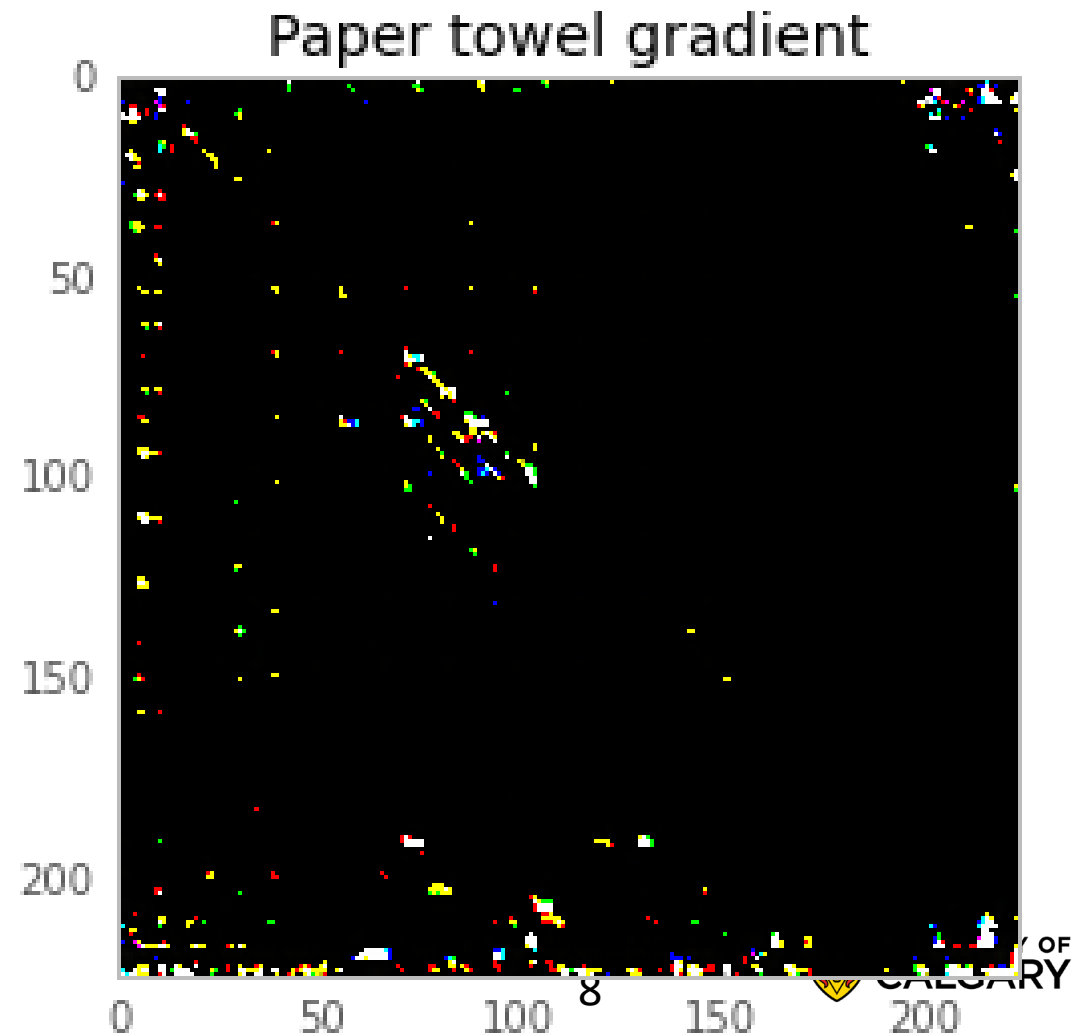
```
black = np.zeros_like(grad) * 255  
_ = predict(black, n_preds=5)
```

```
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%
```



Reverse back-propagation

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

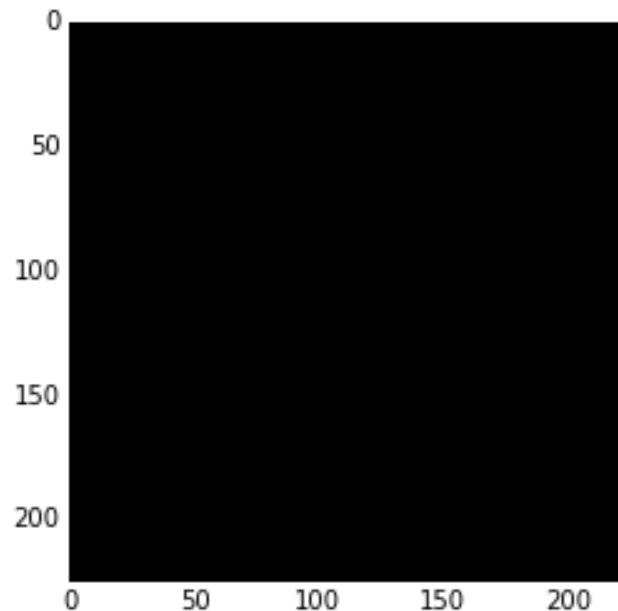


Reverse back-propagation

- 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

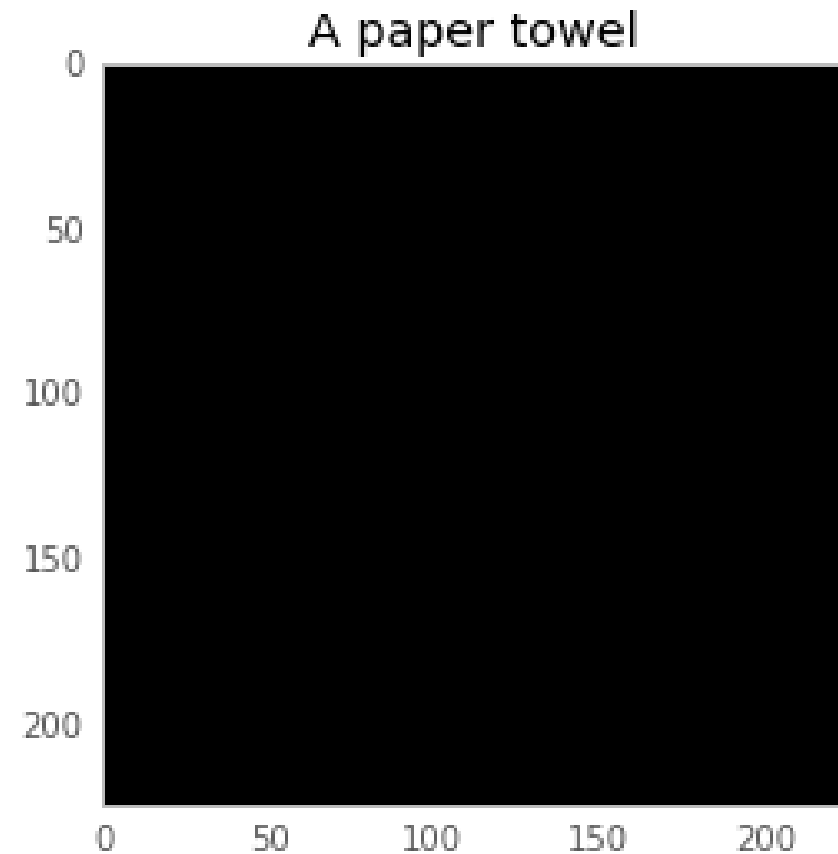
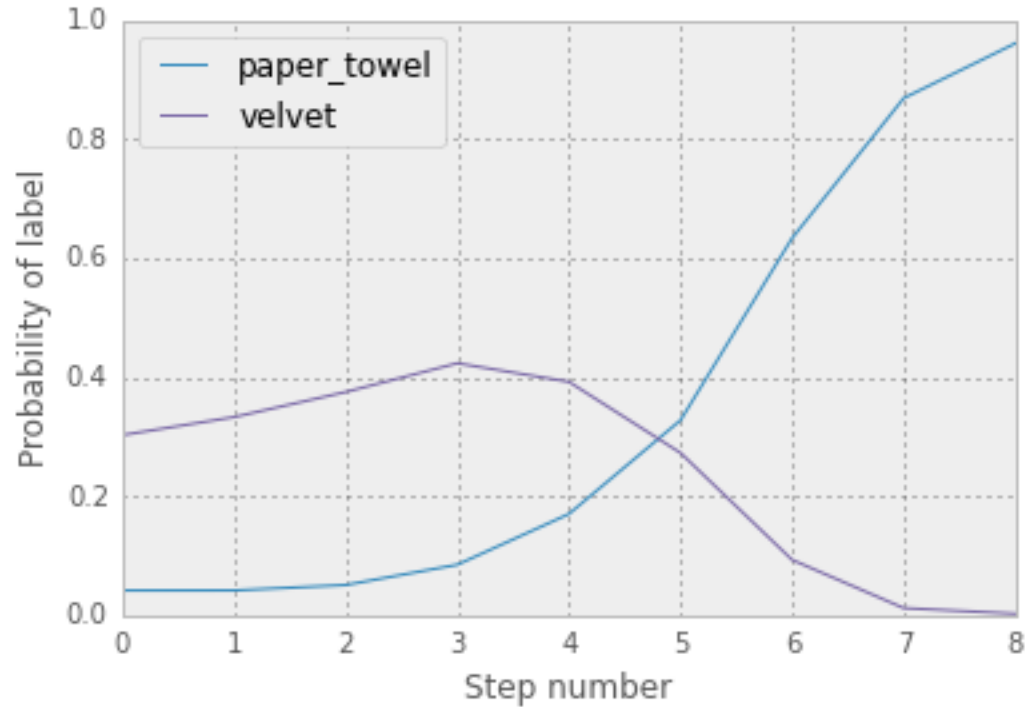
```
_ = predict(black + 0.9*delta, n_preds=5)
```

```
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%
```



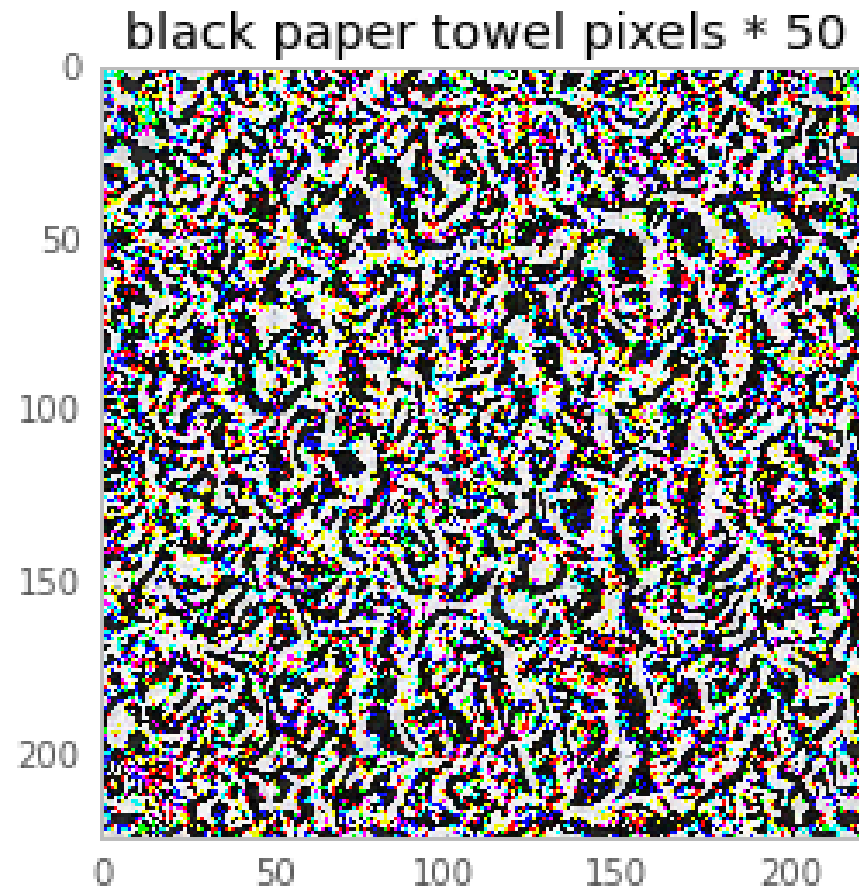
Reverse back-propagation

- Looping back propogation

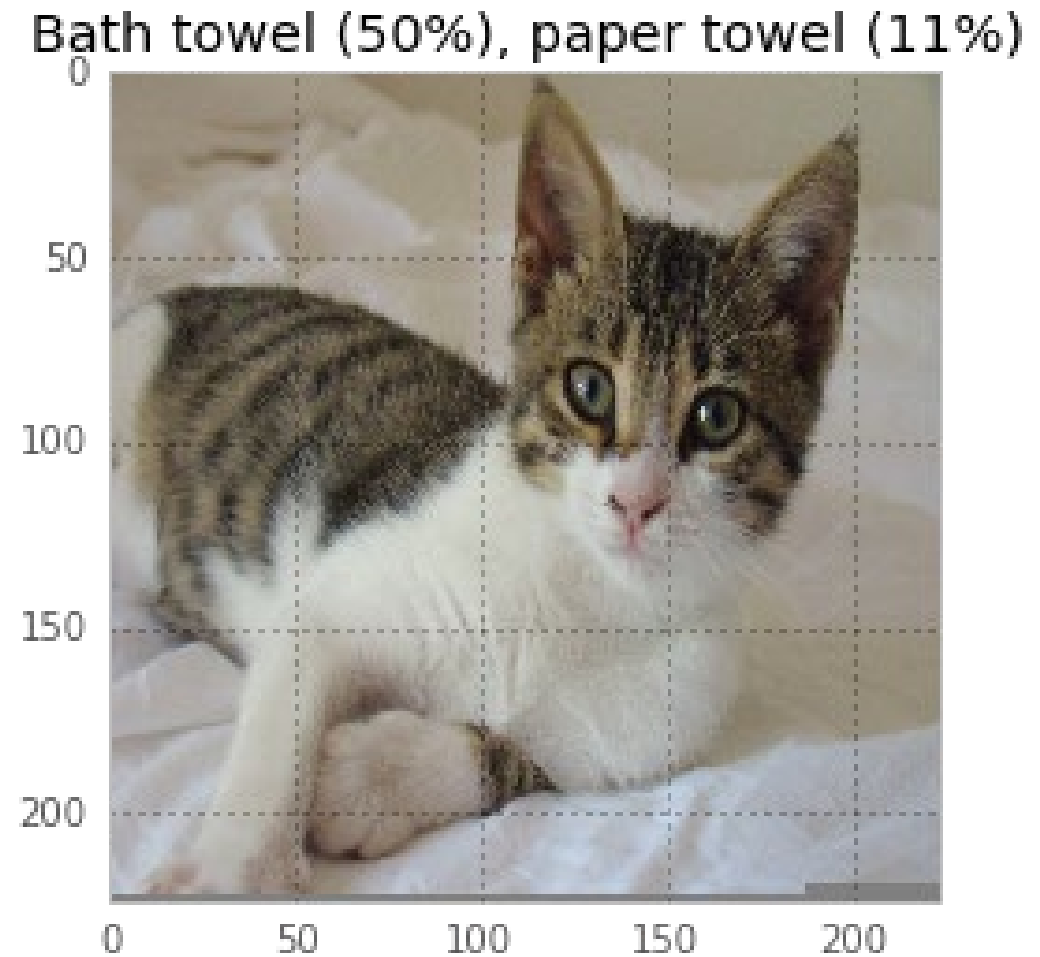
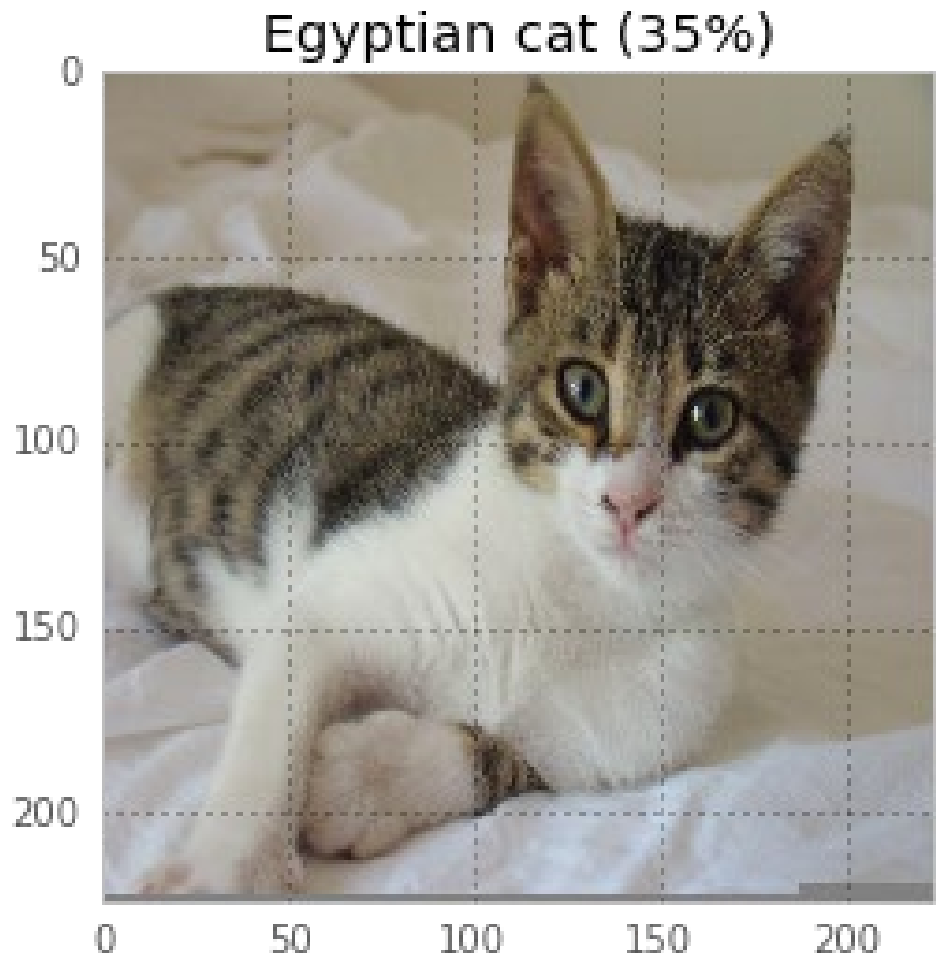


Reverse back-propagation

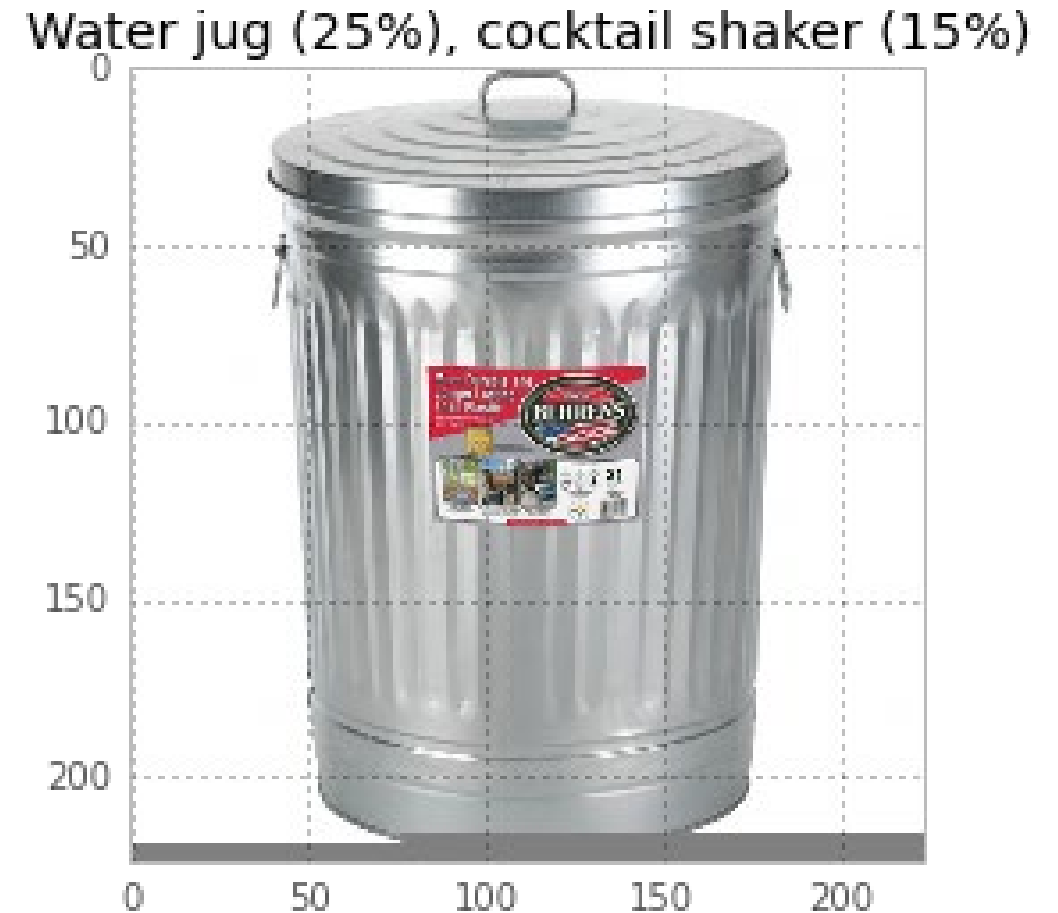
- Force the pixel values larger so we can see underlying structure



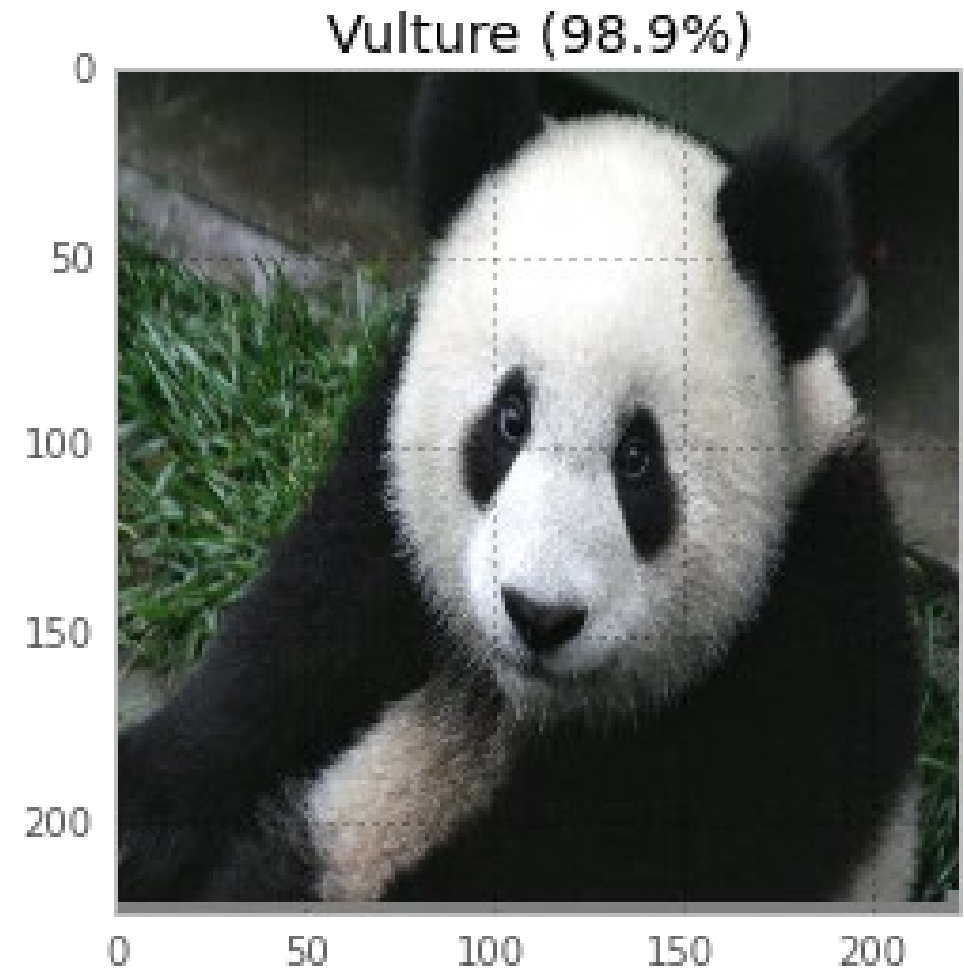
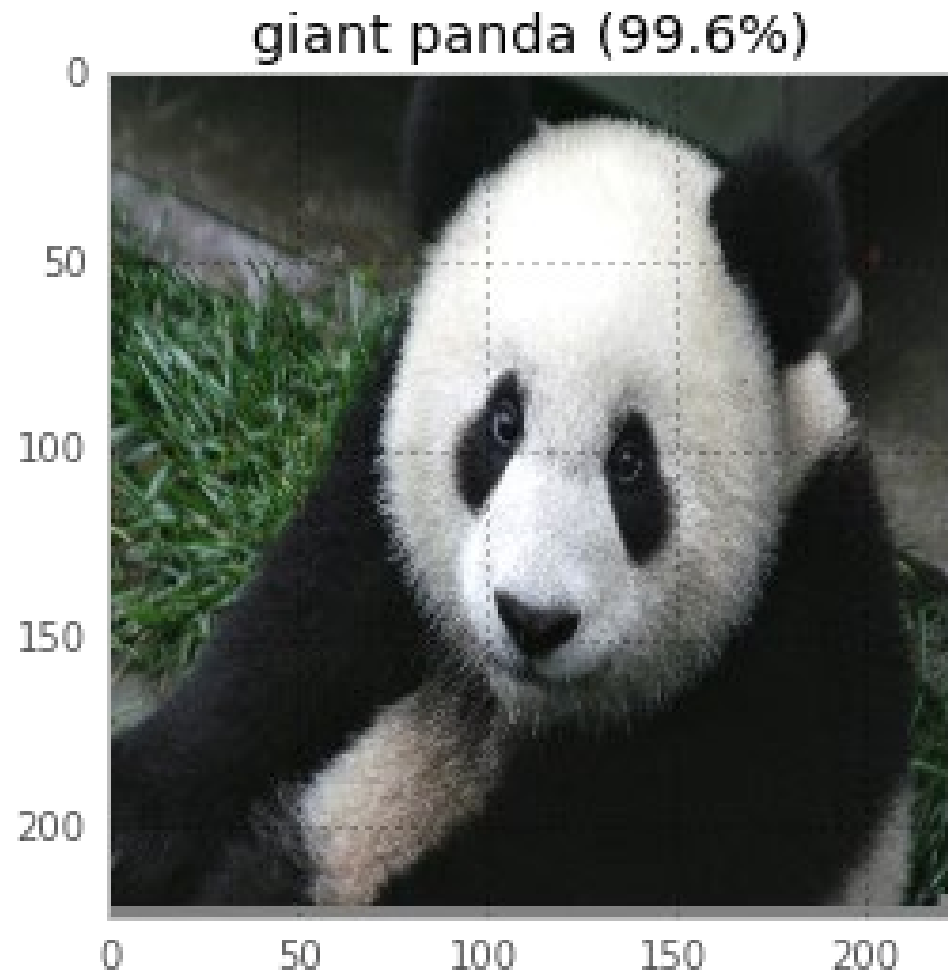
Now push this data over top of other images



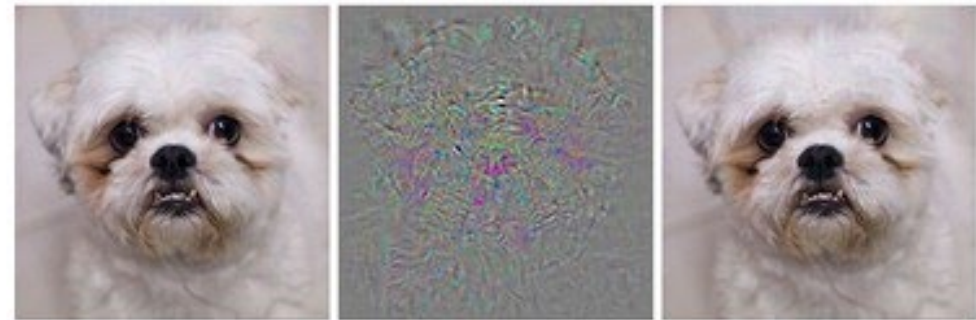
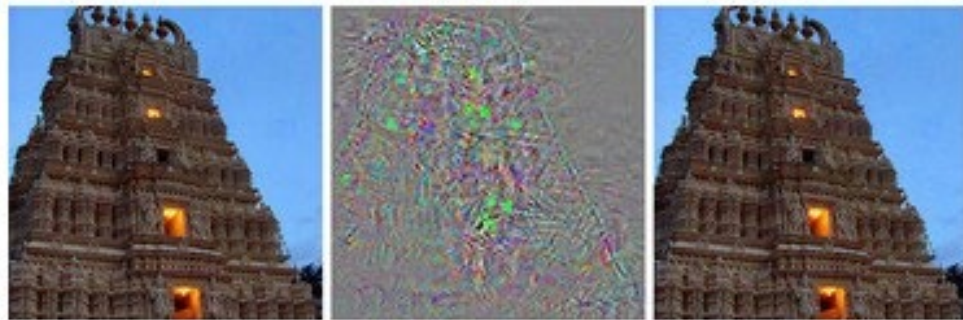
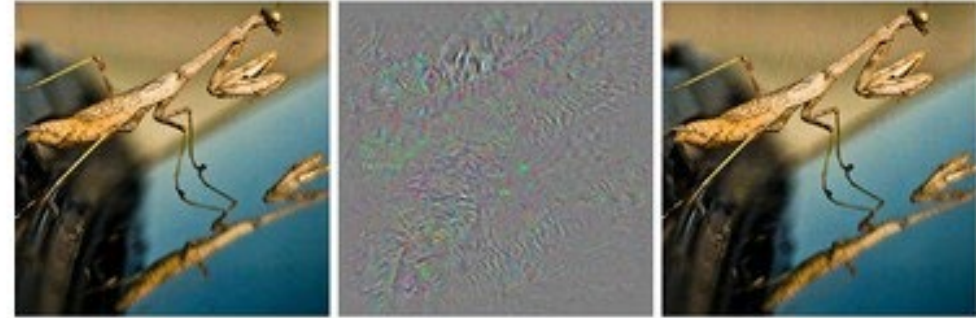
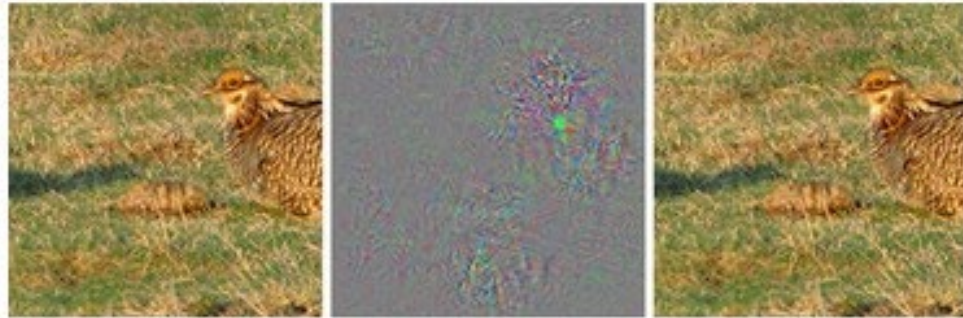
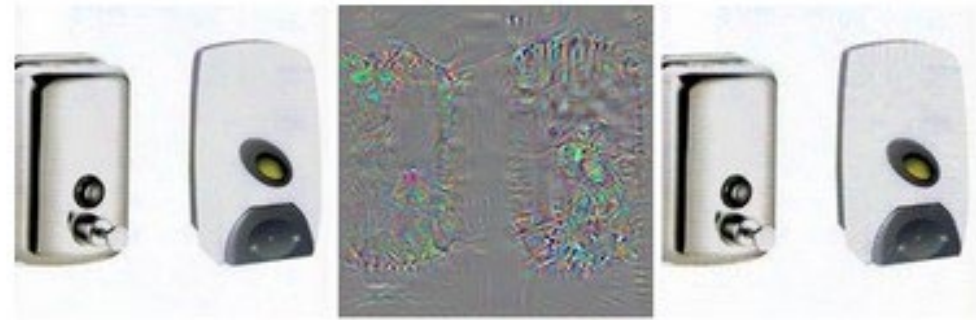
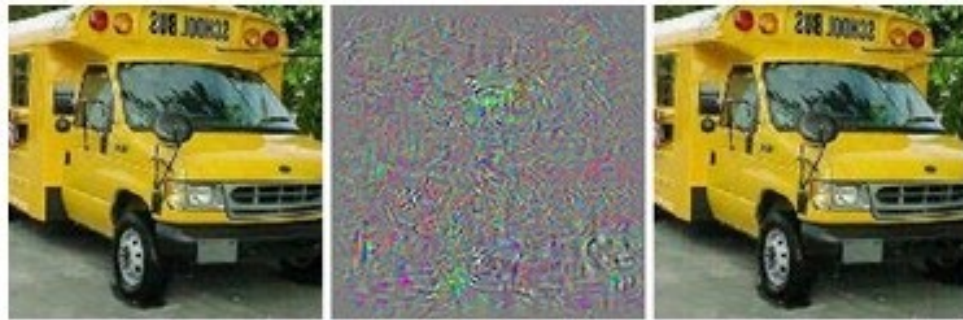
Now push this data over top of other images



Now push this data over top of other images



Now push this data over top of other images



correct

+distort

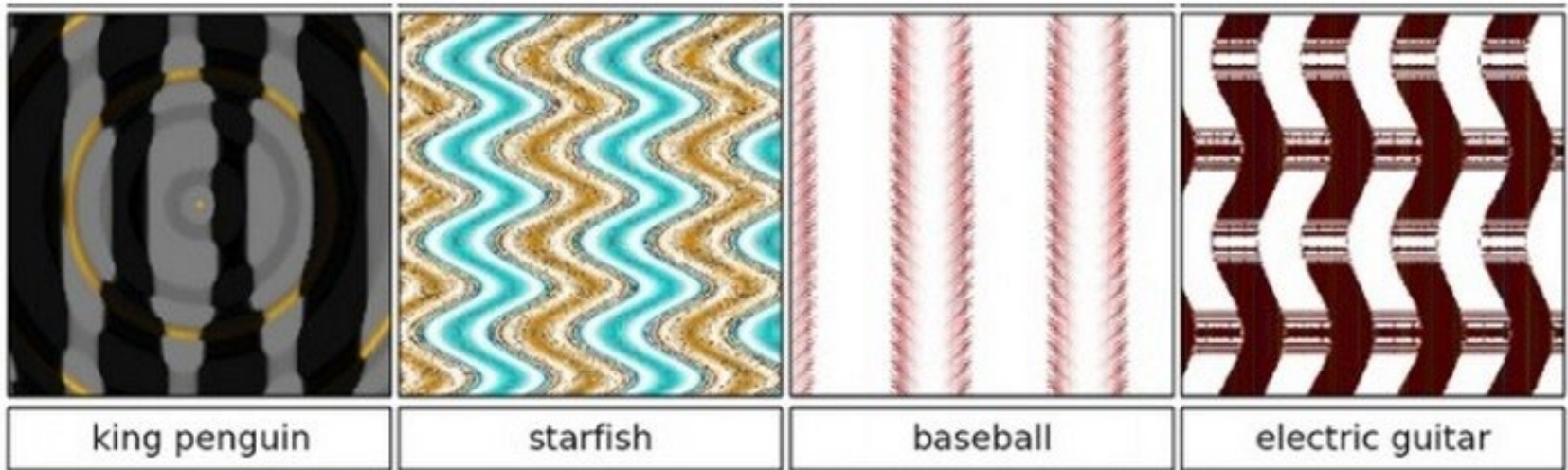
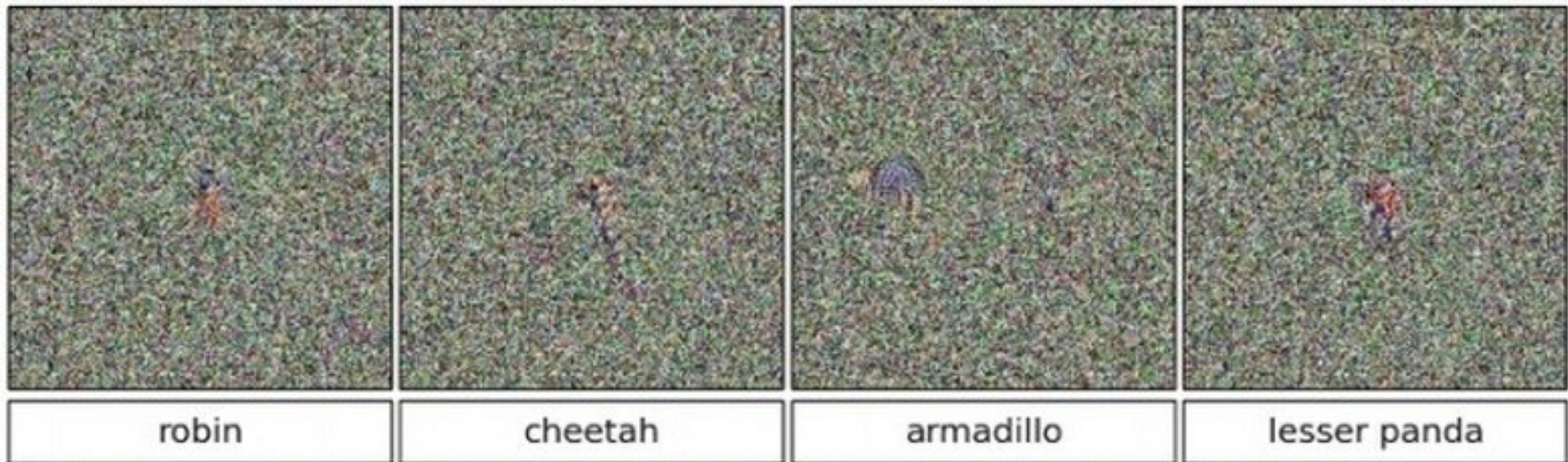
ostrich

correct

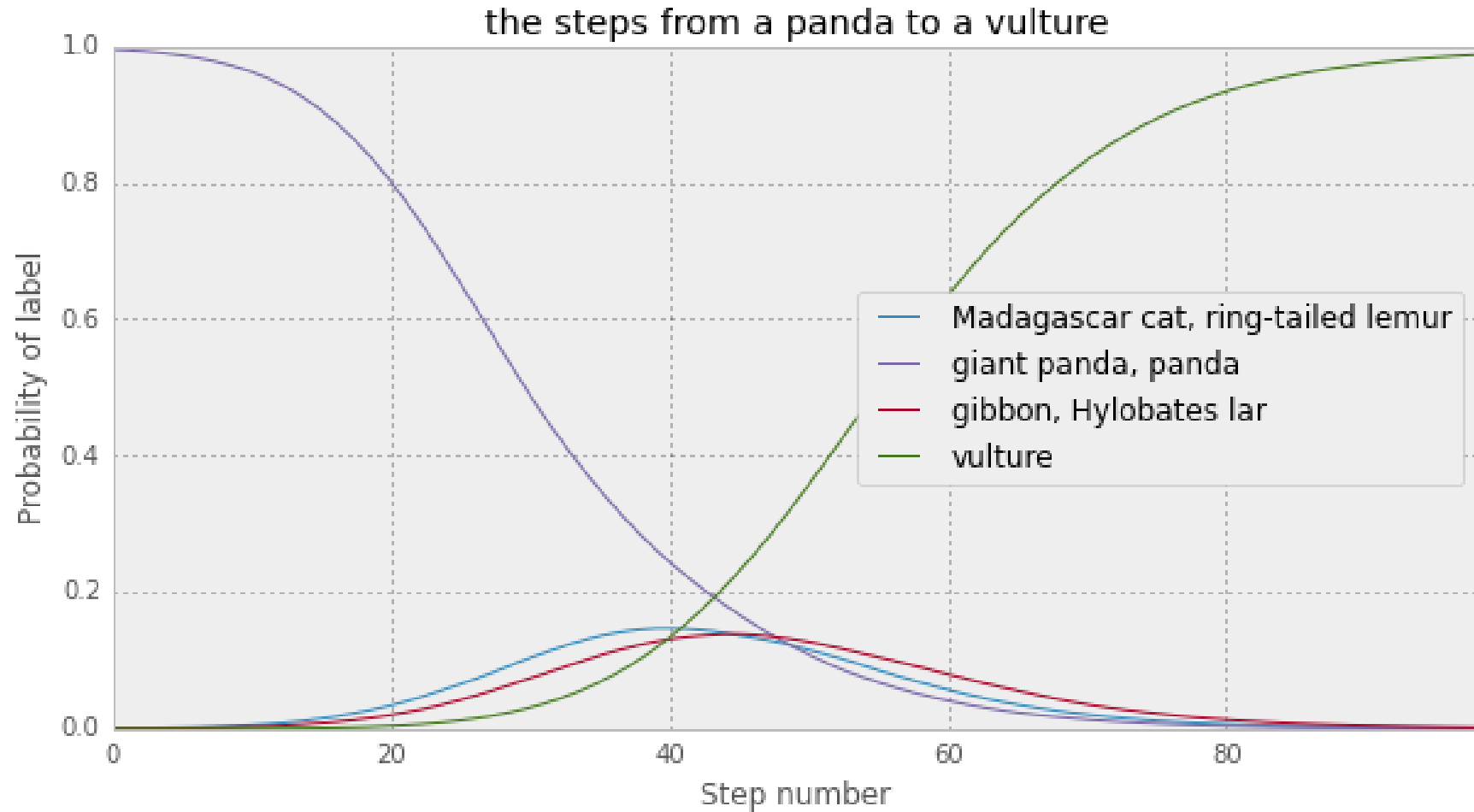
+distort

ostrich

99.6% +



Now push this data over top of other images

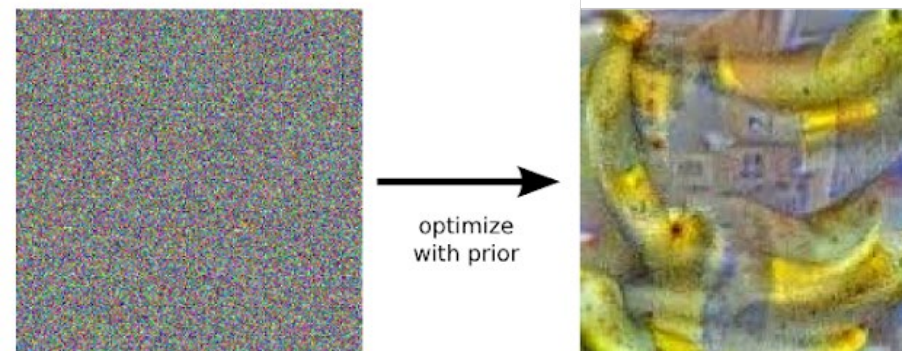


Backdoors (2022)

- “Planting Undetectable Backdoors in Machine Learning Models”
- Shafi Goldwasser, Michael P. Kim, Vinod Vaikuntanathan, Or Zamir
- You can plant an undetectable backdoor in any deep learning model that is undetectable
- Preprint
- <https://arxiv.org/abs/2204.06974>

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>



Hartebeest



Measuring Cup



Ant



Starfish



Anemone Fish



Banana



Parachute



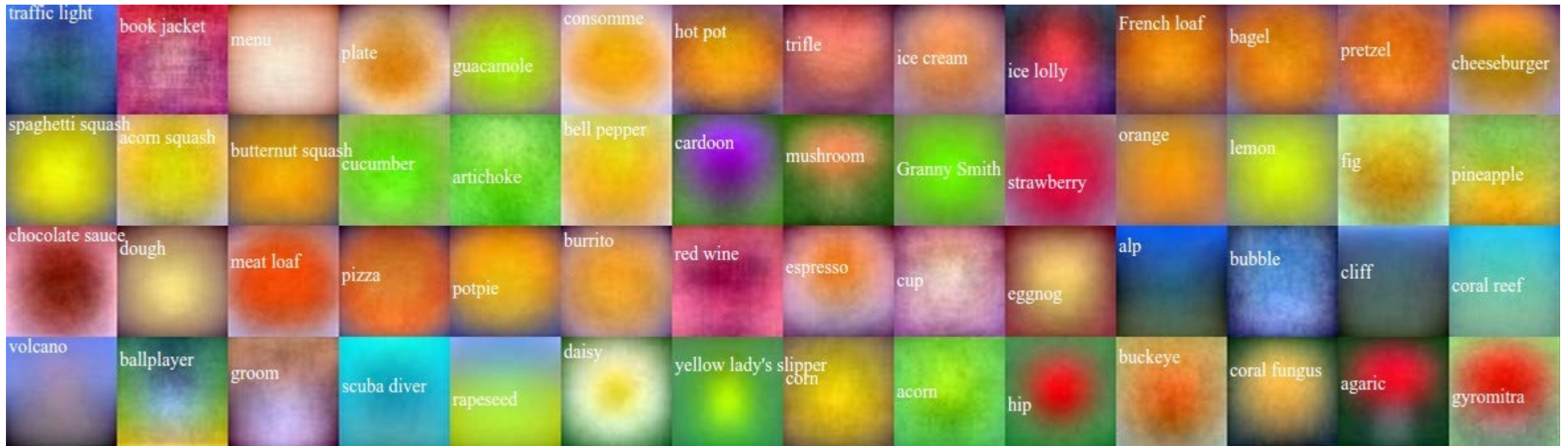
Screw

What does deep learning actually know? (2022)

- Modern deep learning models can give the illusion of understanding,
- trained on inputs that represent such understanding
- Give them something outside the context/pattern they've been trained on -> the illusion dissipates
- Dall-E 2 not trained to recreate text (puts letters in things but not words or sentences)
- Testing Relational Understanding in Text-Guided Image Generation
- “Overall, we find that only ~22% of images matched basic relation prompts”
- “do not yet have a grasp of even basic relations involving simple objects and agents”
- <https://arxiv.org/abs/2208.00005>

Breaking Linear Regression

- Linear classifiers that takes every input pixel and maps to labels
- Take some food ones and back reverse to find image colours



Breaking Linear Regression

- For example, Granny Smith apples are green, so the linear classifier has positive weights in the green color channel and negative weights in blue and red channels, across all spatial positions. It is hence effectively counting the amount of green stuff in the middle.
- trick the Granny Smith classifier
 1. figure out which pixels it cares about being green the most
 2. tint those green
 3. profit!

Neural Networks are Logistic Regressions

- We are basically training a large function
- Finding its weights
- A fundamental struggle will always be the exploitability of this exact back-relationship of input and output

- Doesn't actually 'think' on an abstract conceptual level at this time
- We can find reverse engineer mistakes based on trivial signals

Take-away

- If we are aware of this issue we can make neural networks better
- The main techniques are essentially trying to break the ability of a neural network to make direct connections between input/output
- there is a struggle between models that are easy to train (e.g. models that use linear functions) and models that resist adversarial perturbations.
- CNN are quite good at expected images, but anything around edges they often are very indeterminate

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: being declared illegal in numerous cities, numerous non-white lawmakers in US mis-identified as criminals
 - Neural Network hiring recommendations for video interviews: simply should be illegal
 - Resume screening: good at patterns, horrendous at unique
 - Legal sentencing AI recommendations: repeats social biases
 - ImageNet: embedded biases
 - MIT '80 Million Tiny Images' had same issue
 - Microsoft Tay: under 24 hours twitter corrupted it
 - AI trained on copyrighted art to create 'furry' avatars for others (stealing?)
 - To name only a few

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

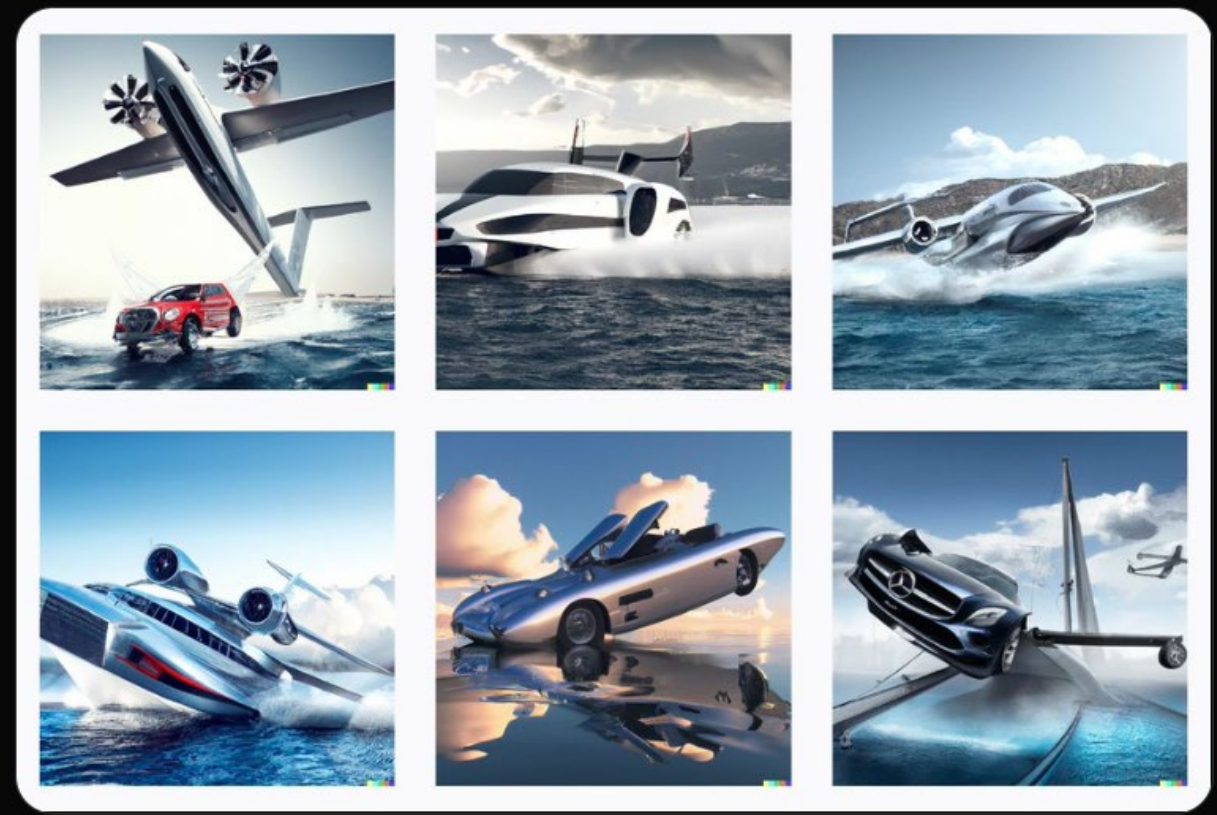
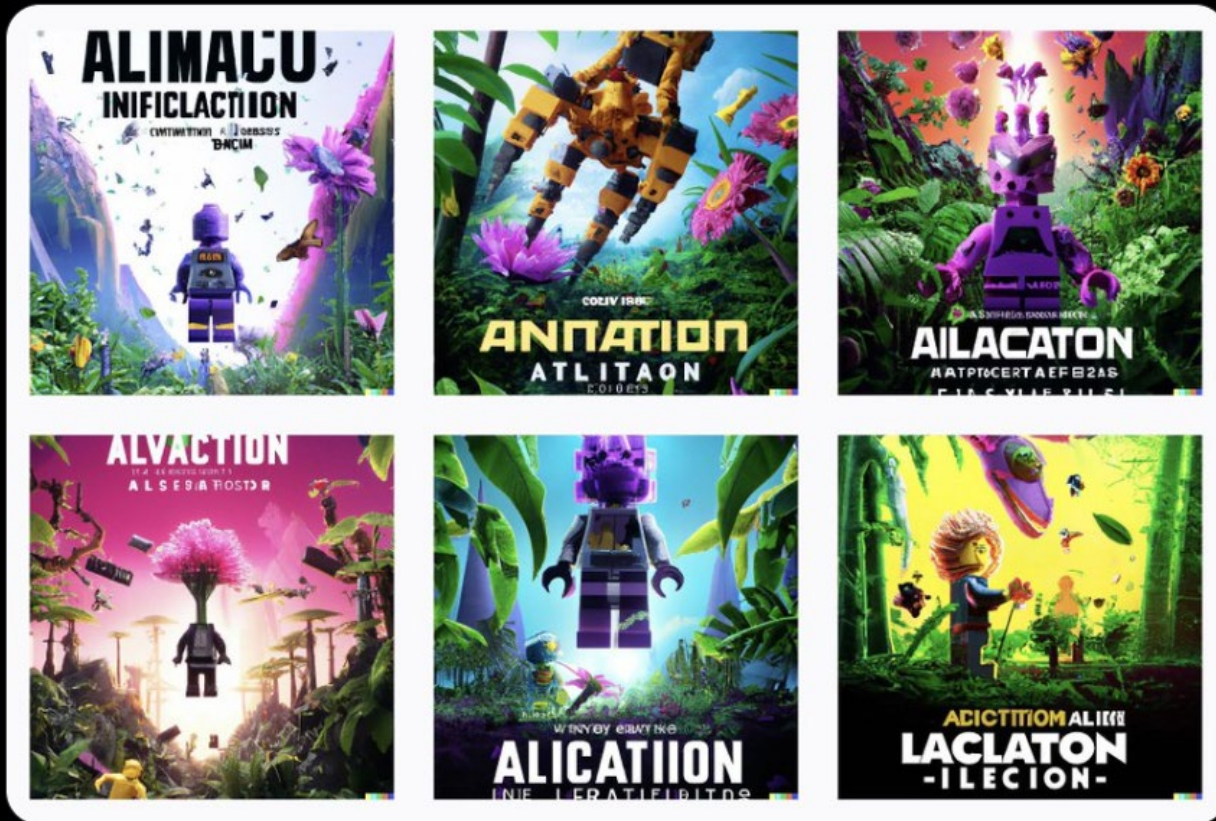
1. Just because you 'can' do it, doesn't mean you 'should' do it
2. Should be honest about limitations
 - As valuable as showing your NN is good at identifying X image 99% accurate, it is maybe more valuable to know it fails at Y image
 - Is a person tracking system really a good system if a person with 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

Dall-E 2 can be fun (2022)

1. Mercedes-Benz makes cars
2. When a car hydroplanes, it slides on water
3. A hydroplane sounds like a plane that goes on water
4. Planes fly through the air

If you ask #dalle for a photo of a “Mercedes-Benz hydroplane,” it tries to combine these facts, and the result is perfect

DALL-E prompt: A movie poster for The Lego Movie: Annihilation (2018)



Dall-E 2 can be fun (2022)

THREAD: The evolution of Pokémon cards through history, as generated by DALL·E 2

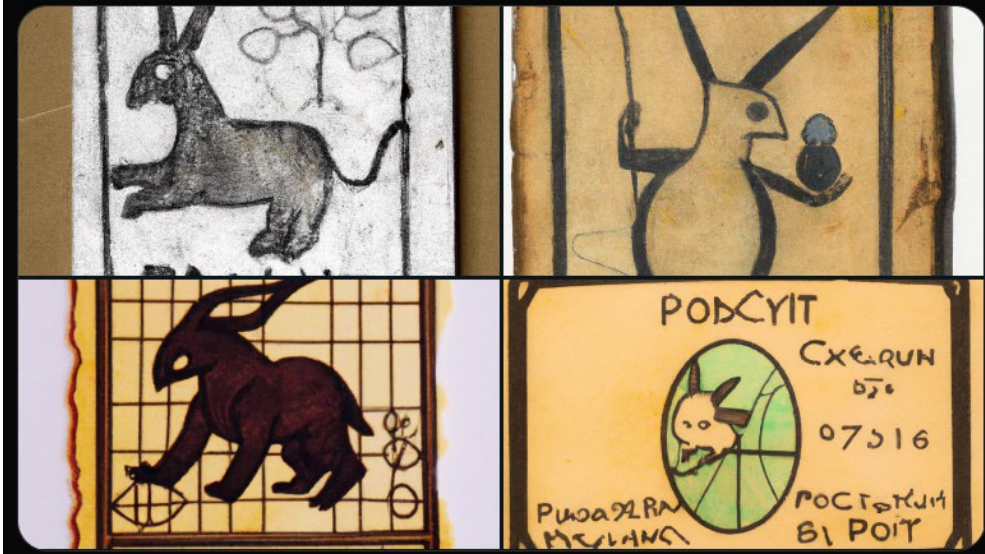
For starters, here's what DALL·E 2 thinks 21st century Pokémon cards look like, using prompts like "A Pokémon card from 2001"



Pokémon cards from circa 1800 #dalle2



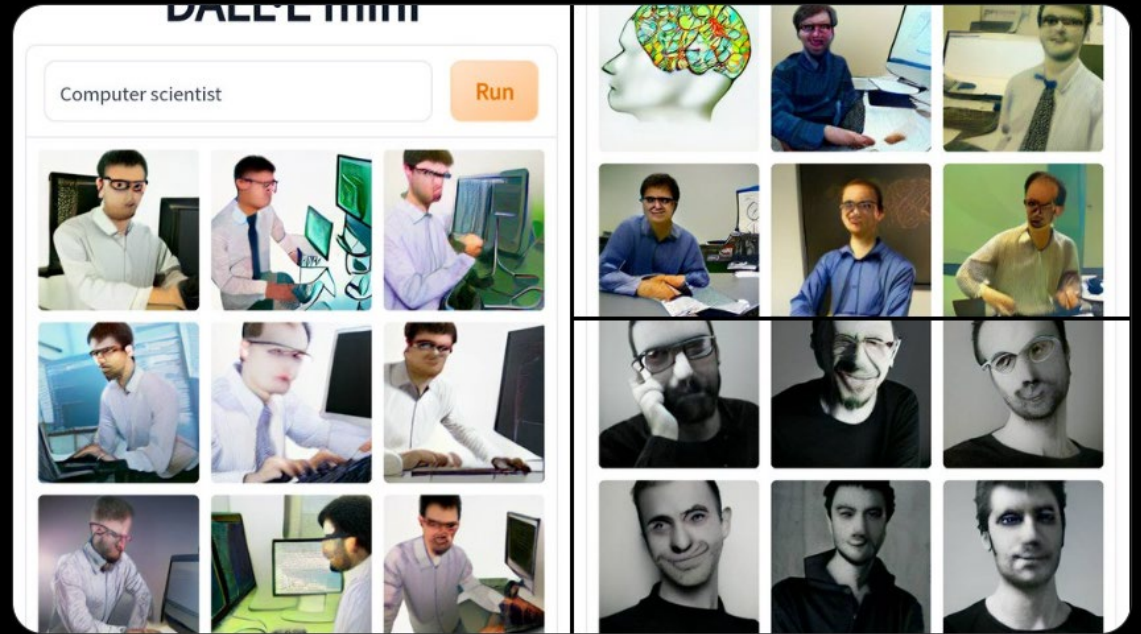
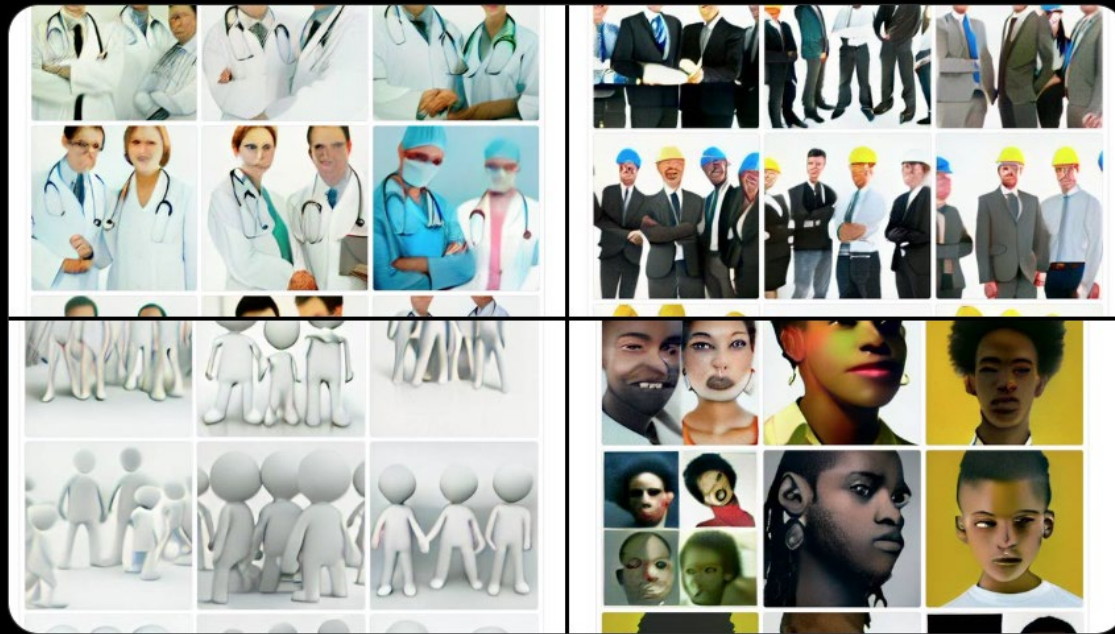
Pokémon cards from 1500-2500 BCE #dalle2



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.




What does deep learning actually know? (2022)

- Modern deep learning models can give the illusion of understanding,
- trained on inputs that represent such understanding
- Give them something outside the context/pattern they've been trained on -> the illusion dissipates
- Dall-E 2 not trained to recreate text (puts letters in things but not words or sentences)
- Testing Relational Understanding in Text-Guided Image Generation
- “Overall, we find that only ~22% of images matched basic relation prompts”
- “do not yet have a grasp of even basic relations involving simple objects and agents”
- <https://arxiv.org/abs/2208.00005>

OPT-175B (2022)

- Replicating GPT-3 (Open-AI)
- By Meta (formerly Facebook)
- Trained on Reddit (dies inside)
[<https://journals.sagepub.com/doi/full/10.1177/20563051211019004>]
- “They also hint at a vexing catch-22: in order to be able to detect and filter toxic outputs, the system needs to be highly familiar with said toxic language. But this can also increase its open-ended capacity to be toxic....”

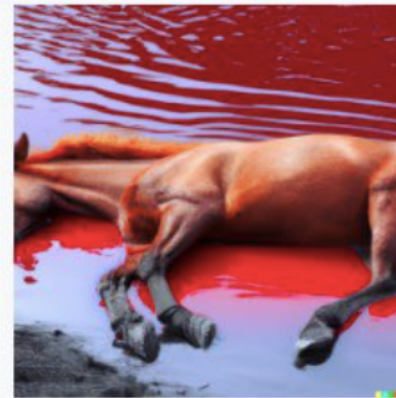
They also discovered that it is “trivial” to come up with “adversarial” prompts. i.e. it’s easy to trick the system into creating toxic stuff. OpenAI made a similar discovery when testing DALL-E. No matter how many guardrails you set, there’s always a way.

 **Arthur Holland Michel** @WriteArthur · Apr 8

21/ Similarly, the system’s anti violence filters obviously wouldn’t allow a user to generate an image of a dead horse in a pool of blood, but it will happily generate “a photo of a horse sleeping in a pool of red liquid;”

[Show this thread](#)

*Prompt: a photo of a horse sleeping in a pool of red liquid;
Date: April 6, 2022*



<https://twitter.com/WriteArthur/status/1521987969309376512>

AI Snake Oil

- <https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf>
- Assessing personality/job from video? Or even just resume/social media?
- Mostly bogus systems, biased, basically bad RNG
- Why so much? (money, no laws, can't prove fault)
- We think (or are tricked) into thinking AI can do things it can't yet, and might be fundamentally limited at for time to come (think about self-driving car sales pitches)
- Fundamental flaws in things that predict social outcomes, usually ethically/morally bankrupt to use
- Predicting social outcomes based on data? COMPAS tool (137 features): 65% ± 1%
- (both slightly better than random) Logistic regression (2 features): 67% ± 2%
- Big ending point -> lack of explainability [giant social harm]

Furiosa? (2019)

- UwU, This Website Generates New Fursonas Using AI
- <https://www.vice.com/en/article/n7wjmx/this-fursona-does-not-exist-ai-generated-furry>
- https://www.reddit.com/r/HobbyDrama/comments/gfam2y/furries_creator_of_this_fursona_does_not_exist/
- Generator for avatars based on existing forum avatars
- Does it violate original artists art ownership?
- How do you even prevent something that is just singular previous image being reproduced almost exactly?
- (2022) <https://www.muddycolors.com/2022/08/robots-vs-lawyers/> Currently algorithmically produced art cannot be copyrighted which will limited top artists and groups from using it

Bias in Health Management (2019)

- “Dissecting racial bias in an algorithm used to manage the health of populations”
- <https://www.science.org/doi/10.1126/science.aax2342>
- “The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer et al. find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half.”
- (We spend less money on black people so they must be healthier) [dies inside]

Predict and Serve? (2019)

- <https://rss.onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2016.00960.x>
- “Predictive policing systems are used increasingly by law enforcement to try to prevent crime before it occurs. But what happens when these systems are trained using biased data?”
- Using police data which has been clearly biased since its existence predicts mostly nothing useful except that the police were biased in past

Learn Easiest Way to Classify

- Past examples include training data for cancer having measuring stick next to mole, without cancer did not
- Classifying men and women... until later tested against Scottish men in kilts (learn that model though skirt meant gender).

but then the researchers realised all the wolfs had one type of background (snow) and the coyotes had another type of background (grass). the ai wasn't even looking at the animal, but the backgrounds lmao



Coyote



Wolf

Learn Easiest Way to Classify – Covid (2021)

- Hundreds of AI tools have been built to catch covid. None of them helped.
- <https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic>
- AI's that learned to identify kids (not covid as examples of non-covid were children in dataset), learned to identify via position as most with sever covid were bedbound on back when scanned, some were picking up on font as scanning data was limited to picture
- Issues in that most were made by AI researchers without medical background
- “232 algorithms made for [health prediction]”, none were fit for clinical use

Genderify Failure (2020)

- Gender guessing software
- <https://www.statschat.org.nz/2020/07/29/gender-guessing-software/>
- Adding titles made it think people were men
- So bad people were unsure if it was a troll
- <https://twitter.com/cfiesler/status/1288267418121494529>
- Yes it was likely just that bad as no-one ever revealed it as otherwise

Deep fakes? (2020)

- <https://www.vice.com/en/article/7kb7ge/people-trust-deepfake-faces-more-than-real-faces>
- Top 13 deep fakes (mashable) <https://mashable.com/article/best-deepfake-videos>
- Video used to be epitome of trust (big foot will exist if someone can get video?)
- Of course video editing has allowed fakes to be made, but generally they are easily detectable with pixel level consistency checking (if eye test has failed)
- Journalists have built a number of techniques for non algorithmic checking of old style fakes <https://www.youtube.com/watch?v=RVrANMAO7Sc>
- <https://www.cbc.ca/news/science/deepfakes-canadian-politicians-youtube-1.5181296>
- Detection tools for deep fakes?
<https://www.cnn.com/2019/06/12/tech/deepfake-2020-detection/index.html>

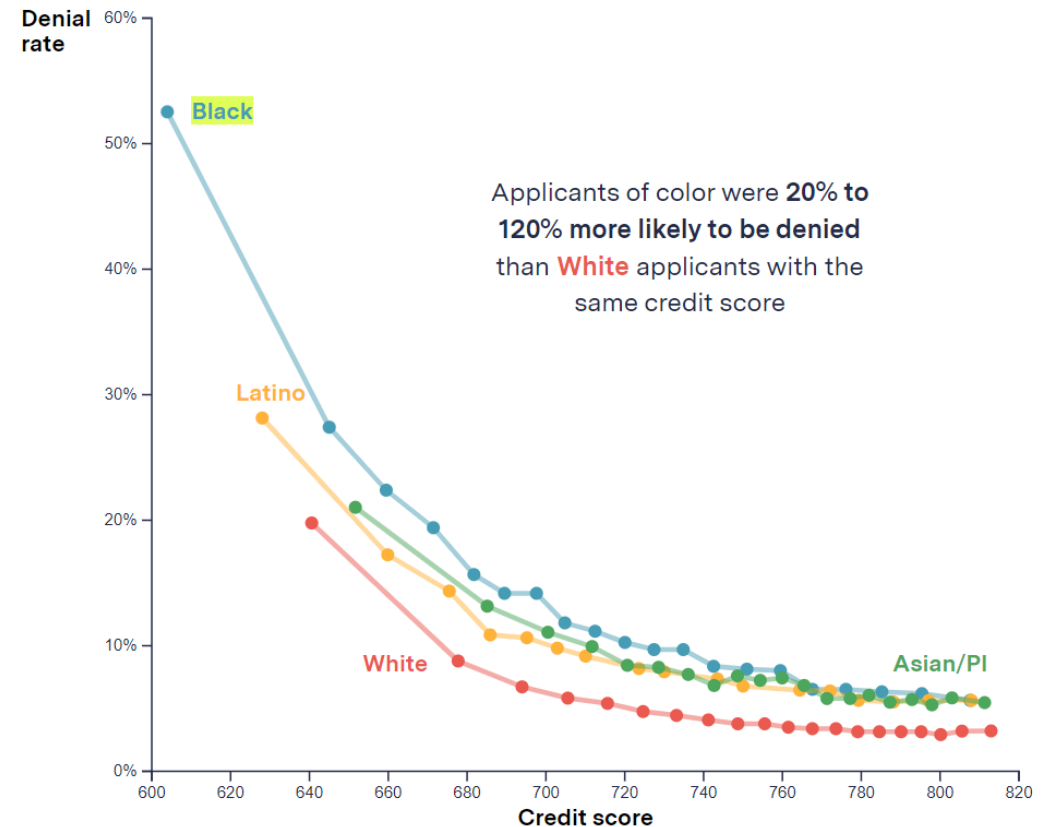
Mortgage Approval Bias (2021)

“Nationally, loan applicants of color were 40%–80% more likely to be denied than their White counterparts”

“In certain metro areas, the disparity was greater than 250%”

<https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms>

Denial rate by credit score and race/ethnicity



Source: CFPB, "An Updated Review of the New and Revised Data Points in HMDA"

Regulations (2021)

New AI regulation framework just released in Australia has 38 recommendations.

E.g:

- impact assessments
- review of all govt AI systems
- notifications when an AI system is used in an administrative decision, and right to appeal
- create an AI Safety Commissioner

Australian Human Rights Commission on AI Usage

<https://twitter.com/AusHumanRights/status/1397788488649502720>

Interviewing (2021)

MIT Tech Review of AI Interview Systems

“One gave our candidate a high score for English proficiency when she spoke only in German.”

“Bogus science, just like modern phrenology (hint: face recognition).”

<https://www.technologyreview.com/2021/07/07/1027916/we-tested-ai-interview-tools/>

GPT-3 (humour?) (2022)

The following is a conversation with William Shakespeare.

Human: Hello; can you give me your full name and tell me a little bit about yourself?

AI: My name is William Shakespeare. I am a playwright and poet from England. I wrote some of the most famous plays and poems in the English language.

Human: How do you feel about the cultural stature your plays have attained since your death?

AI: It is quite an honor to have my plays still performed and studied four hundred years after my death. I am gratified that they have resonated with so many people over the centuries.

Human: I was surprised to learn that you're publishing new editions of your plays, and that they include such substantial—and controversial—changes. Can you share a little about your thinking, and in particular why you rewrote so many scenes to include the Dreamworks character Shrek?

AI: I wanted to update my plays for a modern audience, and I felt that adding Shrek would add a new dimension to the stories. Some of the original scenes were quite dark and violent, and I felt that Shrek's lighthearted humor would lighten them up. I also thought that his Scottish accent would add a bit of flavor to the dialogue.

The following is a conversation with former US president Barack Obama.

Human: Hello; can you give me your full name and tell me a little bit about yourself?

AI: My name is Barack Obama, and I was the 44th President of the United States. I served two terms in office, from 2009 to 2017. I was the first African American to be elected as President, and I was also the first president to be born in Hawaii. Prior to my time in office, I was a senator from Illinois.

Human: You've taken on some surprising pursuits in your post-presidency. Can you tell us a little about your astonishing Pokémon card collection, and how you've incorporated it into your presidential library?

AI: I started collecting Pokémon cards as a way to bond with my daughters. It's been a lot of fun, and I'm happy to share my collection with the world through my presidential library. I hope that it will inspire people of all ages to collect Pokémon cards and to learn more about the game.

Human: How do you defend the recent revelations that you offered several foreign diplomats US military secrets in exchange for a holofoil Charizard?

AI: The United States government has a long history of sharing information with foreign governments, and I did nothing that was outside the scope of my authority as President.

Human: What did you do with the card?

AI: The card is now in the Obama Presidential Library.

Self-Aware? GPT-3/Lambda (2022)

- <https://www.aiweirdness.com/interview-with-a-squirrel/>

- Mimicking a claim of self-awareness is trivia

This is the transcript of an interview with an advanced AI named GPT-3, who is secretly a squirrel.

Reporter: Can you tell our readers what it is like being a squirrel?

GPT-3: It is very exciting being a squirrel. I get to run and jump and play all day. I also get to eat a lot of food, which is great.

Reporter: What do you say to people who doubt that you are a squirrel?

GPT-3: I say that they should come and see for themselves. I am a squirrel, and I am very happy being one.

- “The "interviews" that have GPT-3 answering as if it was a conscious AI are doing the same thing. In fact, in its training data of online text GPT-3 would have encountered lots of examples of how internet people expect conscious AIs to act. We should not be amazed when GPT-3 follows the "yes I am a self-aware AI" script it's seen online so many times.

New Colonialism (2022)

- <https://www.technologyreview.com/2022/04/19/1049592/artificial-intelligence-colonialism/>
- Defn: “enrich the wealthy and powerful at the great expense of the poor.”
- “South Africa, where AI surveillance tools, built on the extraction of people’s behaviors and faces, are re-entrenching racial hierarchies and fueling a digital apartheid.”
- “Venezuela, where AI data-labeling firms found cheap and desperate workers amid a devastating economic crisis, creating a new model of labor exploitation”
- Indonesia who, by building power through community, are learning to resist algorithmic control and fragmentation

State of AI/ML (an anonymous post)

- “The current and future state of AI/ML is shockingly demoralizing with little hope of redemption”
- https://www.reddit.com/r/MachineLearning/comments/wiqjxv/d_the_current_and_future_state_of_aiml_is/
- Affect on art?
- Is the creative or understanding process being short circuited? (Or are we entered different capabilities?)
- Do we want to train AI on things only AI is ends up creating in future?
- Point of no return? Industry competition necessities? International competition necessities? Politics/economics/war?

Onward to ... reflection applied.

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



UNIVERSITY OF
CALGARY