MAS.S63 Final Project

An attempt at understanding the differences between the features used by deep convolutional neural nets and those used by humans for image identification

Keeley Erhardt

Motivating Questions

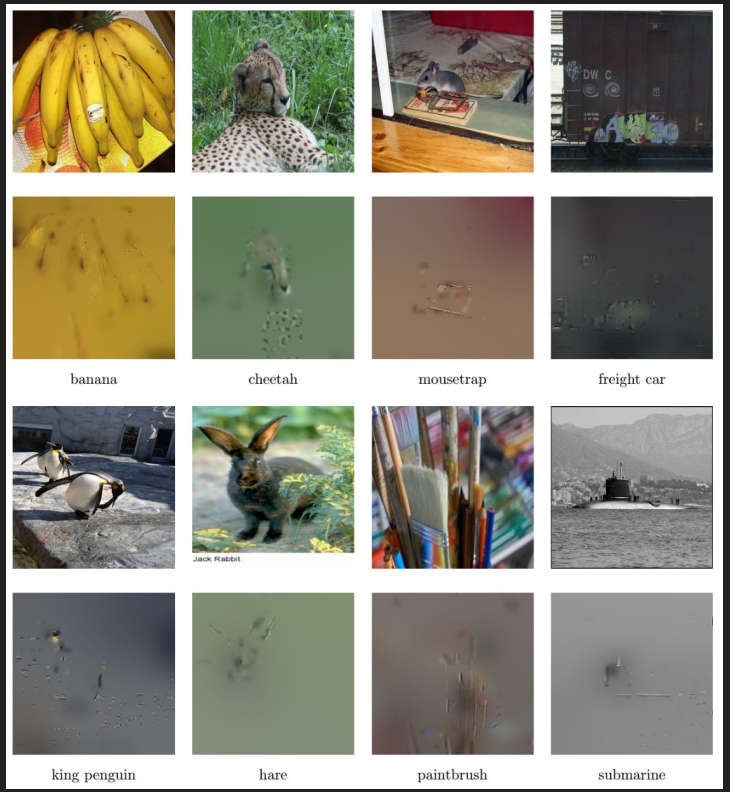
1. Do the identification mechanisms used by humans and DCNNs converge for certain image classes?

2. Is there a relationship between these image classes and the labels assigned to fooling images?

Deep Neural Networks are Easily Fooled (Nguyen et al. 2014)

- A fooling image is an image that state-of-the-art neural nets classify with a high degree of confidence as one thing, whereas a human observer would find the image unrecognizable

Exploring DCNN Features via Signal-Energy Reduction (Kraft 2016)



Signal-Reduction Algorithm

(a) Starting Image

(b) Starting Laplacian Pyramid

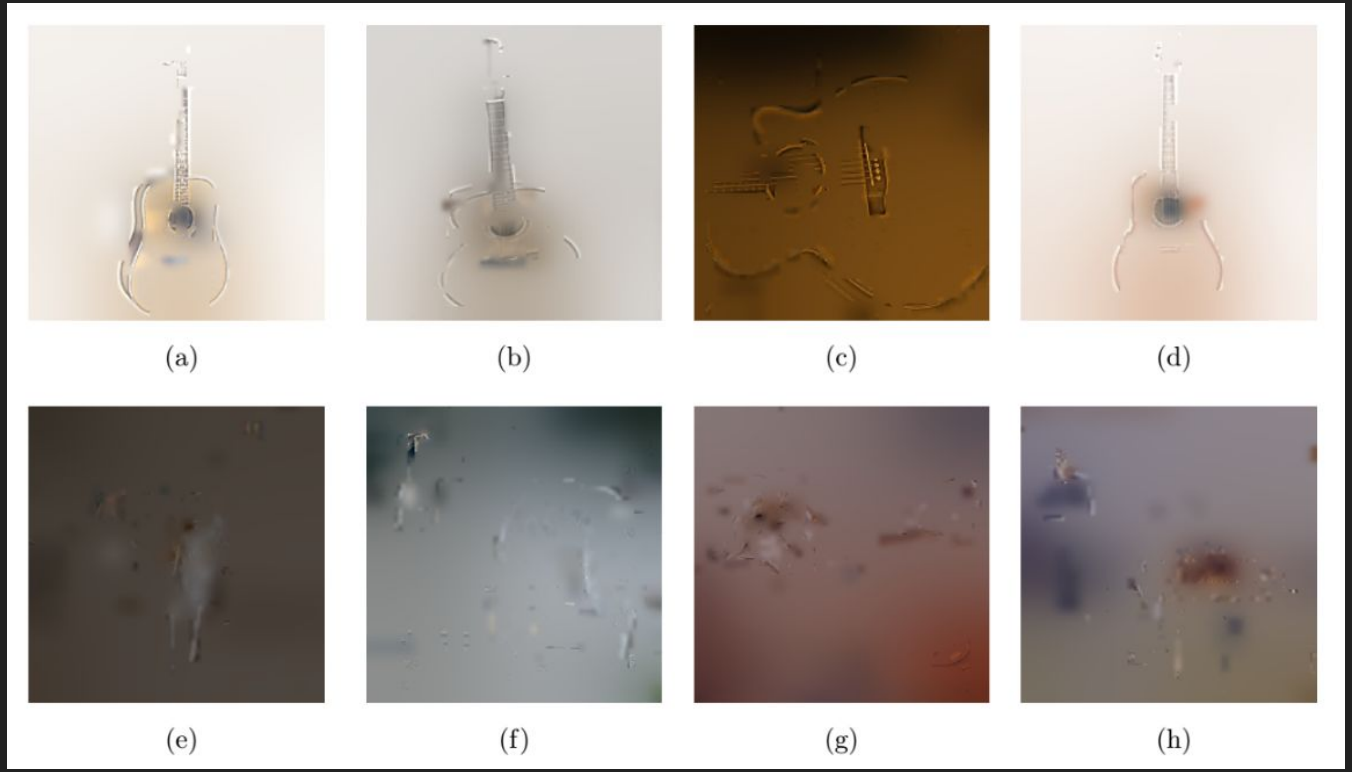
(c) Final Laplacian Pyramid

(d) Final Image

Experiment Idea

Image	Code	Label	Fooling Image Human Recognizable
	n03000134	chainlink fence	Yes
	n03729826	matchstick	Yes
	n03942813	ping-pong ball	Yes
	n04356056	sunglasses, dark glasses, shades	Yes
	n02056570	king penguin, Aptenodytes patagonica	No
	n02317335	starfish, sea star	No
	n03272010	electric guitar	No
	n04074963	remote control, remote	No

Acoustic Guitar + Russian Wolfhound Minimal-Energy Images

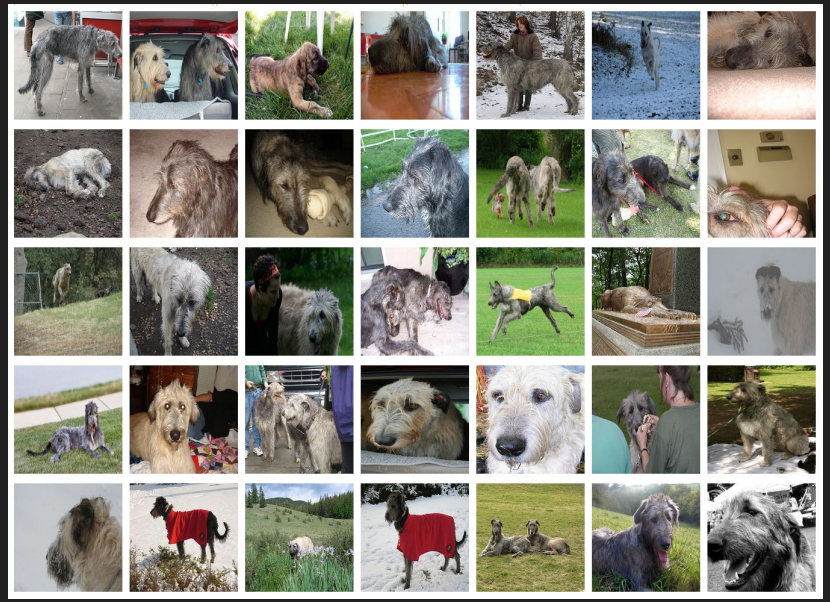


Two Explanations

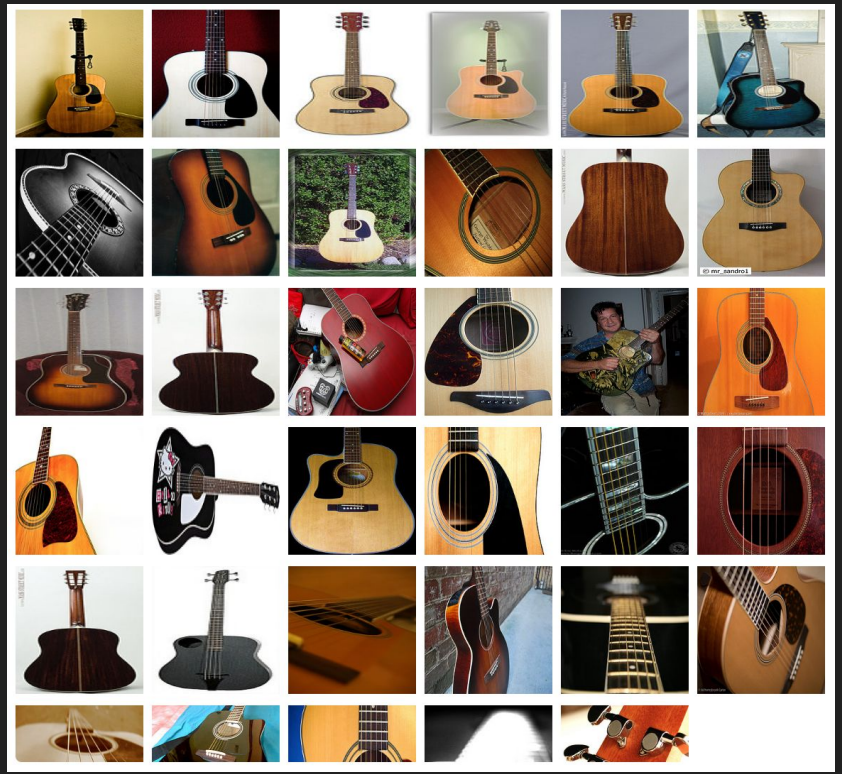
1. The network is tuned to identify many specific types of dogs so it has more units dedicated to this image type than others (i.e. it is less overfit so therefore harder to fool)

2. There are so many dog classes that the EA has trouble finding an image that scores high in a specific category but low in related categories

Russian Wolfhounds versus Irish Wolfhounds

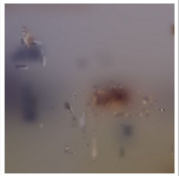


Acoustic Guitar Synset



Labelling of Minimal-Energy Images: Russian Wolfhound

Image	Tensor Flow (AlexNet)	Tensor Flow (Inception-v3)
	1. Arctic fox, white fox, Alopex lagopus 0.128337 2. white wolf, Arctic wolf, Canis lupus tundrarum 0.11411 3. **borzoi, Russian wolfhound 0.0553836** 4. ice bear, polar bear, Ursus Maritimus, Thalarctos maritimus 0.0475548 5. Samoyed, Samoyede 0.0361909	1. barn spider, Araneus cavaticus 0.25208 2. nematode, nematode worm, roundworm 0.20965 3. isopod 0.02492 4. tick 0.02034 5. ant, emmet, pismire 0.01542
	1. tub, vat 0.0637832 2. shower curtain 0.062885 3. bathtub, bathing tub, bath, tub 0.0543703 4. nematode, nematode worm, roundworm 0.0378981 5. beaker 0.0377563	1. nematode, nematode worm, roundworm 0.32972 2. barn spider, Araneus cavaticus 0.19129 3. isopod 0.06626 4. Petri dish 0.03621 5. common newt, Triturus vulgaris 0.03092

	1. ice bear, polar bear, Ursus Maritimus, Thalarctos maritimus 0.208913 2. Arctic fox, white fox, Alopex lagopus 0.166923 3. great white shark, white shark, man-eater, man-eating shark, Carcharodon carcharias 0.0965338 4. hammerhead, hammerhead shark 0.0665496 5. stingray 0.0581107	1. candle, taper, wax light (score = 0.50401) 2. consomme 0.06374 3. eggnog 0.04933 4. West Highland white terrier 0.04152 5. seashore, coast, seacoast, sea-coast 0.01940
	1. axolotl, mud puppy, Ambystoma mexicanum 0.0520174 2. tick 0.0315102 3. barn spider, Araneus cavaticus 0.0313211 4. ant, emmet, pismire 0.02819 5. isopod 0.0258127	1. consomme (score = 0.09062) 2. Petri dish (score = 0.08425) 3. isopod (score = 0.06651) 4. nematode, nematode worm, roundworm (score = 0.06197) 5. goldfish, Carassius auratus (score = 0.04198)

Labelling of Minimal-Energy Images: Acoustic Guitar

Image	Tensor Flow (AlexNet)
	1. **acoustic guitar 0.544734** 2. hook, claw 0.0659736 3. plunger, plumber's helper 0.0309352 4. banjo 0.0278781 5. electric guitar 0.0257582
	1. plunger, plumber's helper 0.467679 2. **acoustic guitar 0.109538** 3. switch, electric switch, electrical switch 0.0274923 4. ladle 0.025571 5. electric guitar 0.0218776
	1. **acoustic guitar 0.449543** 2. electric guitar 0.155197 3. joystick 0.146449 4. plunger, plumber's helper 0.0944864 5. soap dispenser 0.0224255
	1. electric guitar 0.0864191 2. beaker 0.0752944 3. nematode, nematode worm, roundworm 0.0469728 4. mouse, computer mouse 0.045615 5. **acoustic guitar 0.0439272**

Motivating Questions

1. Do the identification mechanisms used by humans and DCNNs converge for certain image classes?

2. Is there a relationship between these image classes and the labels assigned to fooling images?

Appendix

- Inceptionism: Going Deeper into Neural Networks, Mordvintsev, Olah, and Tyka (2015)
- Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images, Nguyen et al. (2015)
- Understanding Deep Image Representations by Inverting Them, Mahendran and Vedaldi (2014)
- Inverting Visual Representations with Convolutional Networks, Dosovitskiy and Brox (2016)
- Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, Simonyan, Vedaldi, and Zisserman (2014)
- Object Recognition with Informative Features and Linear Classification, Vidal-Naquet and Ullman (2003)
- Exploring DCNN Features via Signal-Energy Reduction, Kraft (2016)