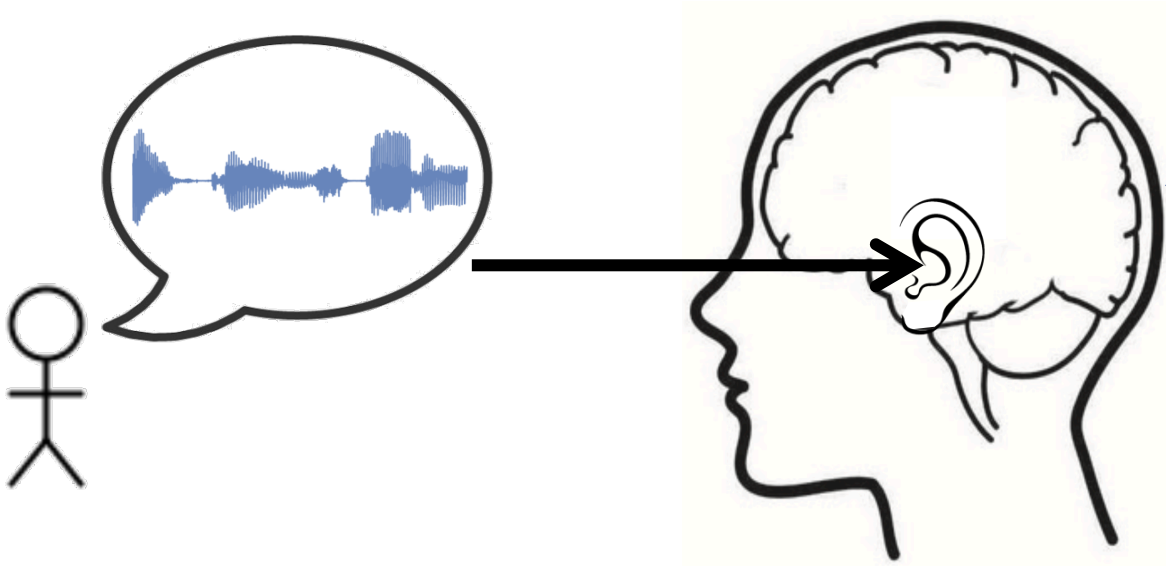
Metamers of neural networks reveal divergence from human perceptual systems

Jenelle Feather

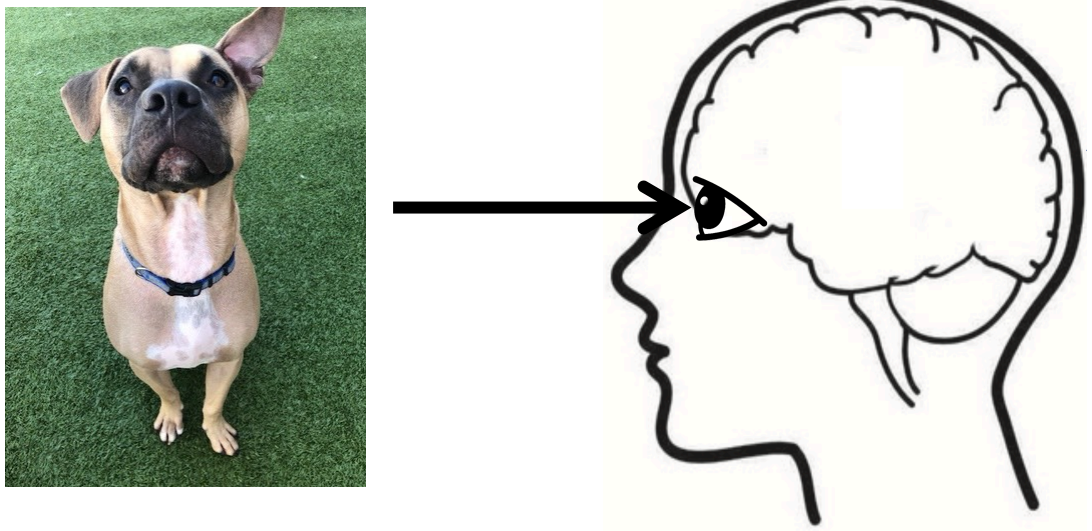
Lab for Computational Audition @ MIT

CSGF Program Review 2020



What you hear:

`I got a new pet!`



What you see:

DOG

Real Neurons inspired McCulloch & Pitts Neuron (1942)

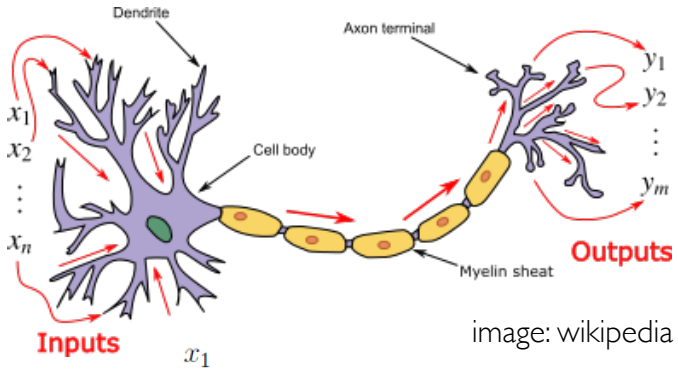
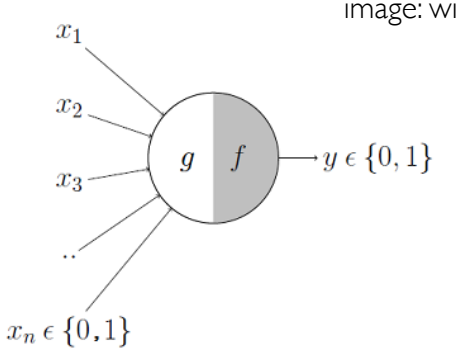
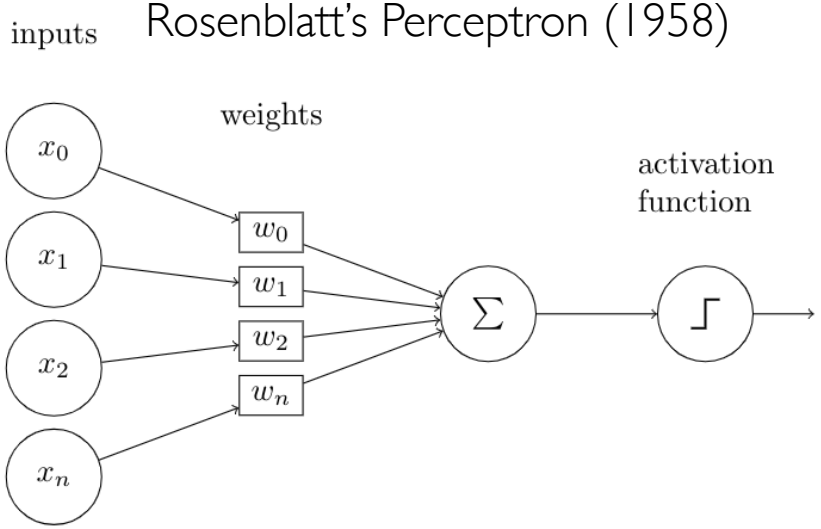


image: wikipedia



Rosenblatt's Perceptron (1958)



Electronic 'Brain' Teaches Itself

The Navy last week demonstrated the embryo of an electronic computer named the Perceptron which, when completed in about a year, is expected to be the first non-living mechanism able to "perceive, recognize and identify its surroundings without human training or control." Navy officers demonstrating a preliminary form of the device in Washington said they hesitated to call it a machine because it is so much like a "human being without life."

Dr. Frank Rosenblatt, research psychologist at the Cornell Aeronautical Laboratory, Inc., Buffalo, N. Y., designer of the Perceptron, conducted the demonstration. The machine, he said, would be the first electronic device to think as the human brain. Like humans, Perceptron will make mistakes at first, "but it will grow wiser as it gains experience," he said.

The first Perceptron, to cost about $100,000, will have about 1,000 electronic "association cells" receiving electrical impulses from an eyelike scanning device with 400 photocells. The human brain has ten billion responsive cells, including 100,000,000 connections with the eye.

Difference Recognized

The concept of the Perceptron was demonstrated on the Weather Bureau's $2,000,000 IBM 704 computer. In one experiment, the 704 computer was shown 100 squares situated at random either on the left or the right side of a field. In 100 trials, it was able to "say" correctly ninety-seven times whether a square was situated on the right or left. Dr. Rosenblatt said that after having seen only thirty to forty squares the device had learned to recognize the difference between right and left, almost the way a child learns.

When fully developed, the Perceptron will be designed to remember images and information it has perceived itself, whereas ordinary computers remember only what is fed into them on punch cards or magnetic tape.

Later Perceptrons, Dr. Rosenblatt said, will be able to recognize people and call out their names. Printed pages, longhand letters and even speech commands are within its reach. Only one more step of development, a difficult step, he said, is needed for the device to hear speech in one language and instantly translate it to speech or writing in another language.

Self-Reproduction

In principle, Dr. Rosenblatt said, it would be possible to build Perceptrons that could reproduce themselves on an assembly line and which would be "conscious" of their existence.

Perceptron, it was pointed out, needs no "priming." It is not necessary to introduce it to surroundings and circumstances, record the data involved and then store them for future comparison as is the case with present "mechanical brains." It literally teaches itself to recognize objects the first time it encounters them. It uses a camera-eye lens to scan objects or survey situations, and an electrical impulse system, patterned point-by-point after the human brain does the interpreting.

The Navy said it would use the principle to build the first Perceptron "thinking machines" that will be able to read or write.

Real Neurons inspired McCulloch & Pitts Neuron (1942)

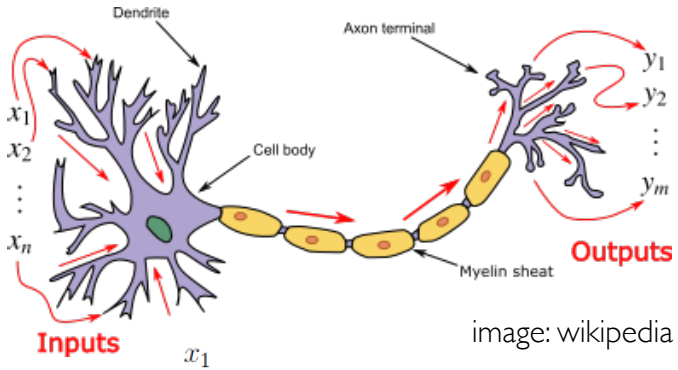
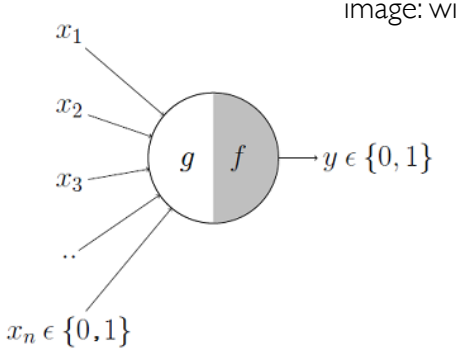
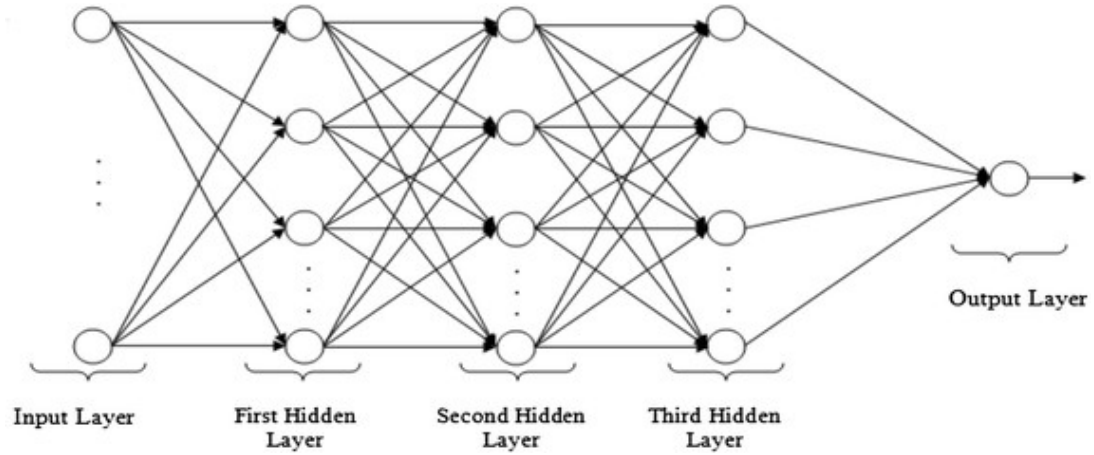


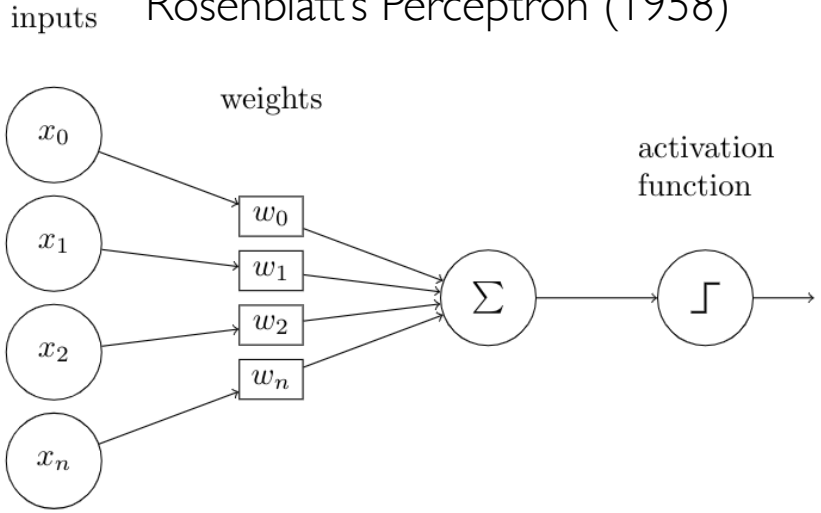
image: wikipedia



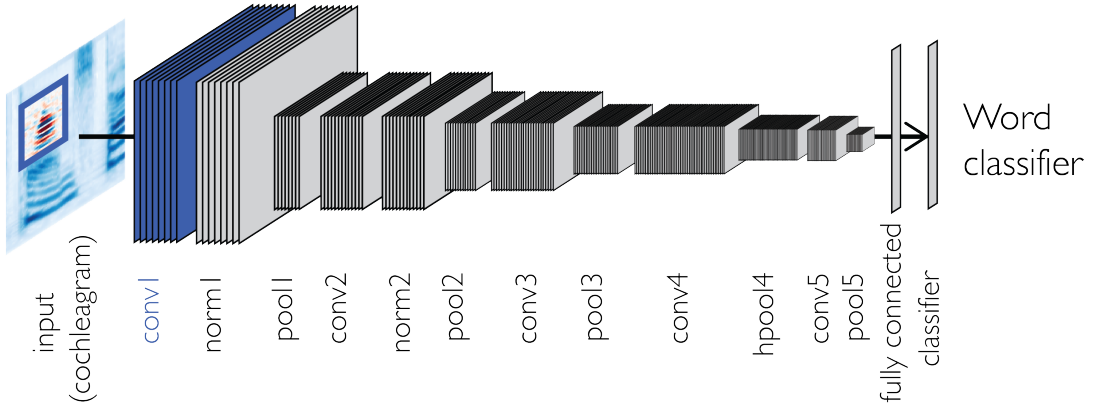
Multi-Layer Perceptrons (1970s/1980s)

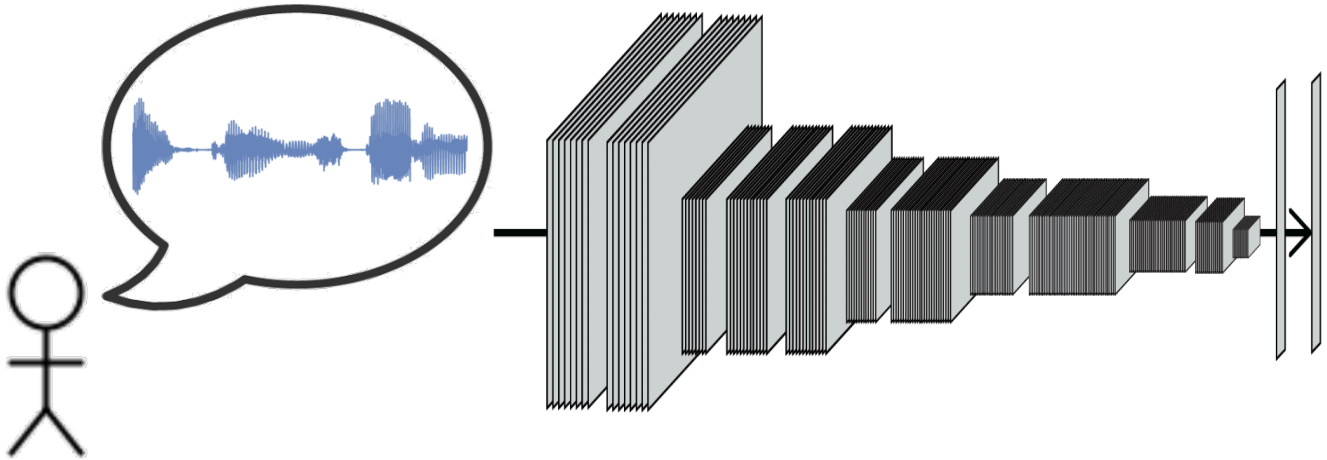


Rosenblatt's Perceptron (1958)



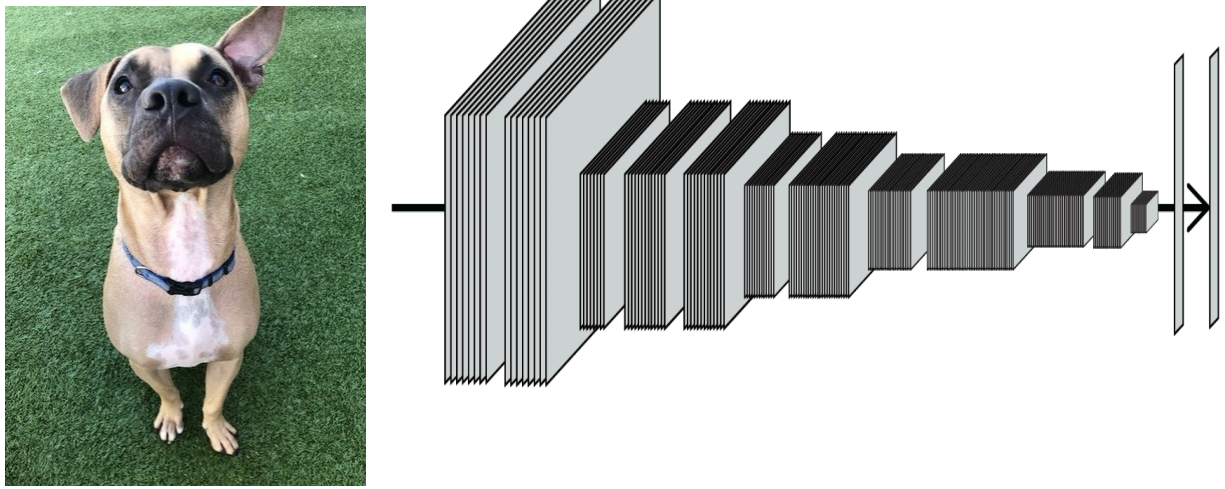
Deep Convolutional Neural Networks (2011-)





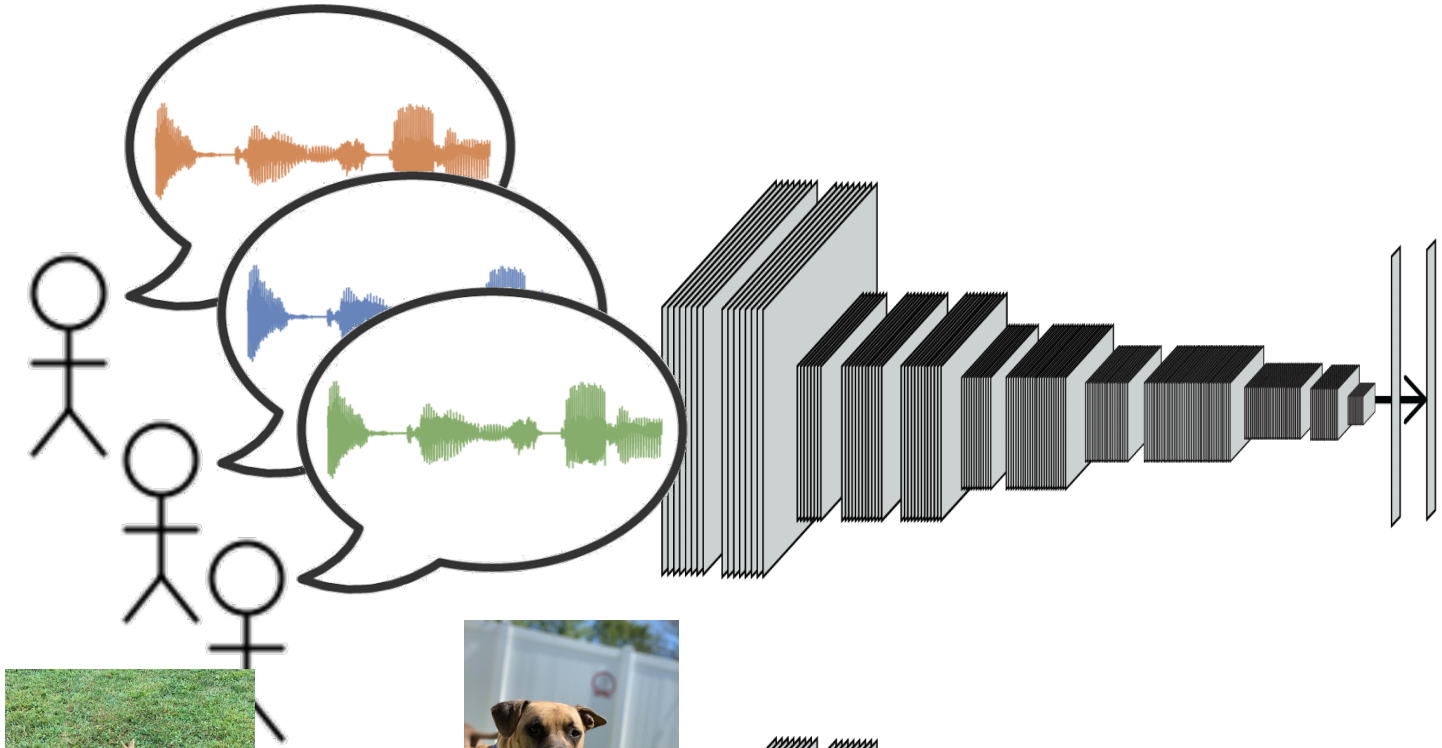
PREDICTED TEXT:

`I got a new pet!`



OBJECT CLASS:

DOG



PREDICTED TEXT:

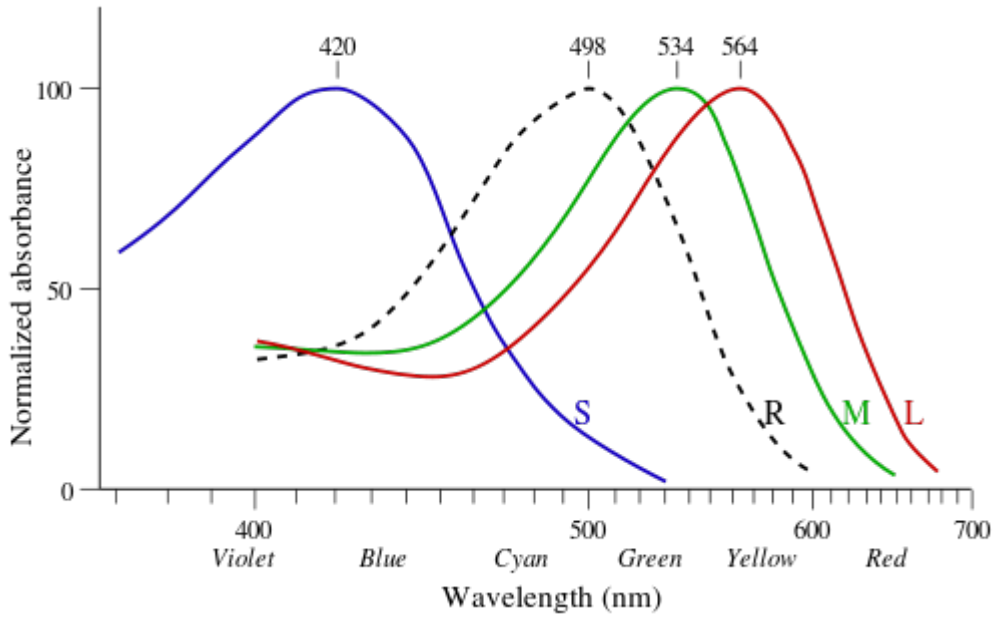
I got a new pet!

OBJECT CLASS:

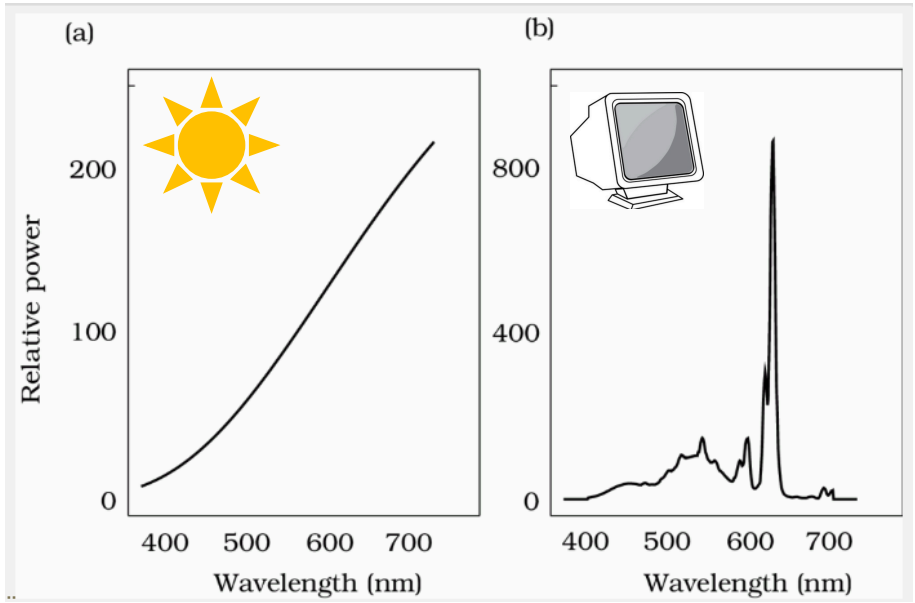
DOG

Metamer: two stimuli that are physically different, but produce the same responses within a system

Beginnings of metamerism: Human Color Vision



r = Bx **r` = Bx`**



Foundations of Vision, Wandell (1995)

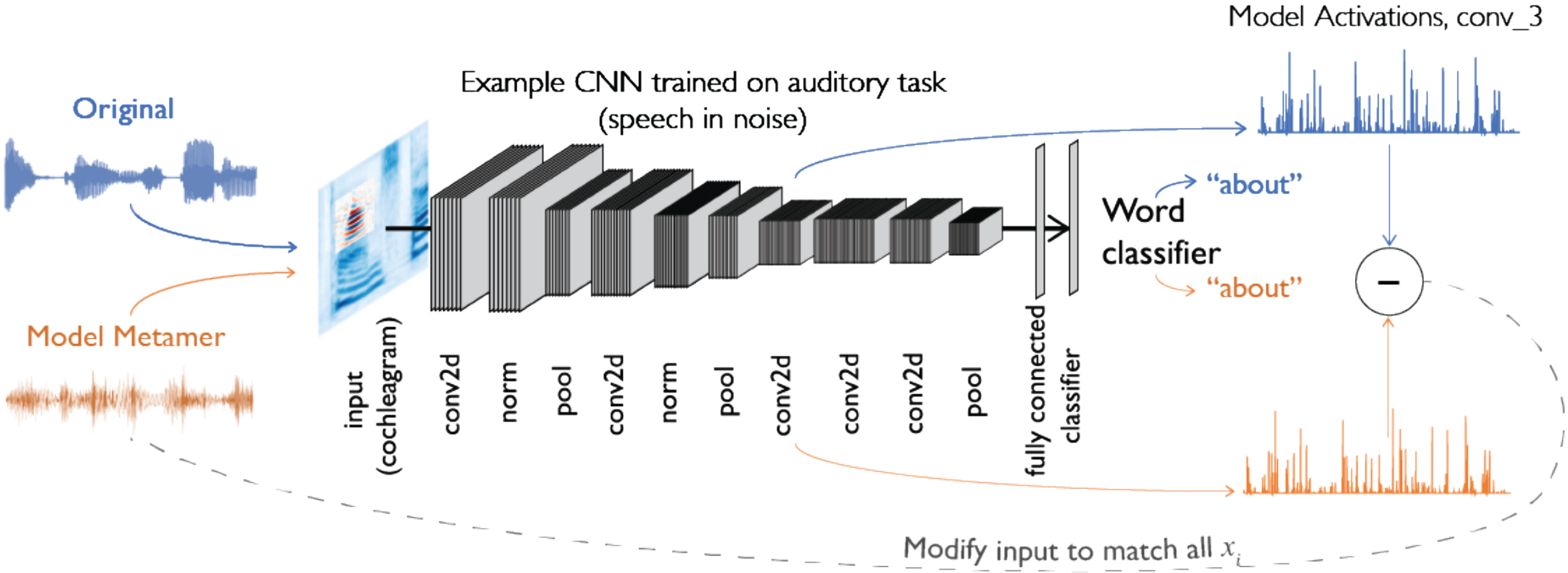
$$\text{Cone absorptions} \begin{pmatrix} L \\ M \\ S \end{pmatrix} = \begin{pmatrix} \text{L cone wavelength sensitivity} \\ \text{M cone wavelength sensitivity} \\ \text{S cone wavelength sensitivity} \end{pmatrix} \begin{pmatrix} \text{Test spectral} \\ \text{power distribution} \end{pmatrix}$$

r **B** **x**

When **r = r`** the two light sources will be perceived as the same color

Main idea: A good model of human perception will share invariances (and thus metamers) with humans

Main idea: A good model of human perception will share invariances (and thus metamers) with humans



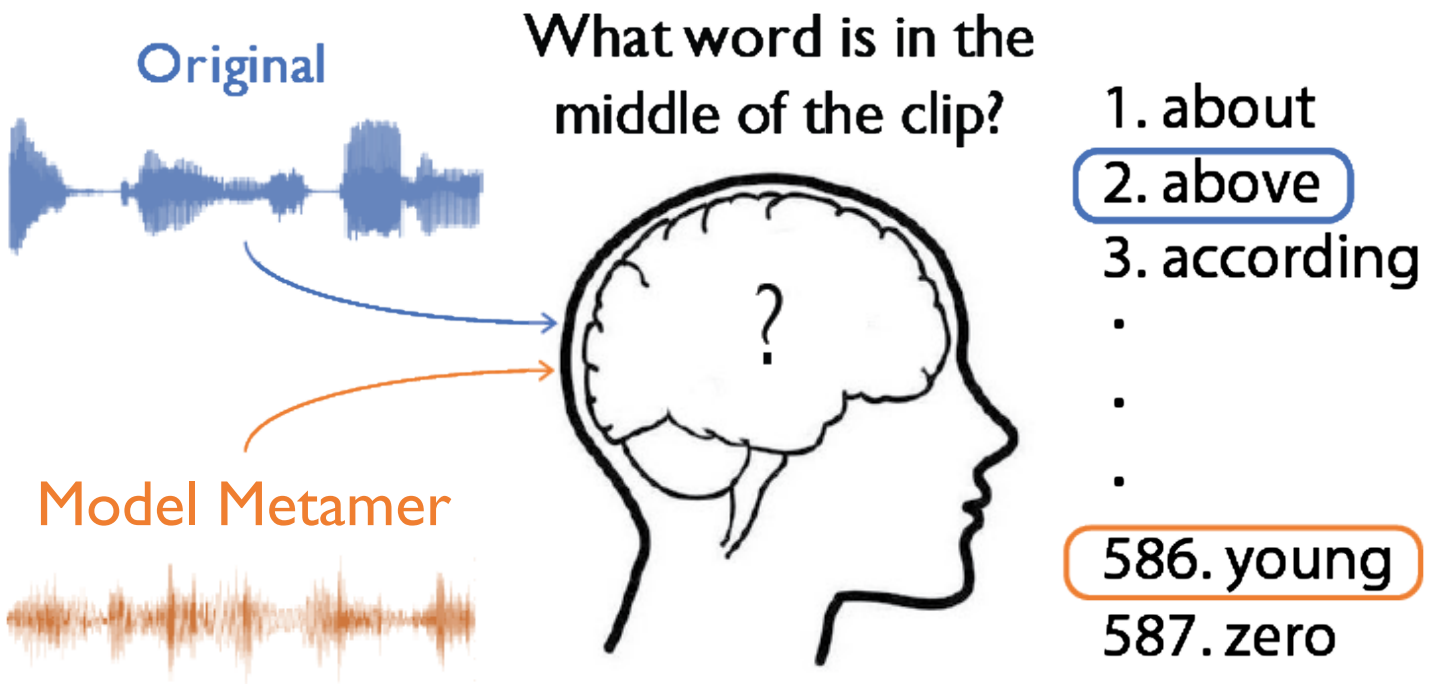
Feather et al., NeurIPS (2019)

Natural question: Are model metamers metameric for humans?

We evaluate with a recognition test

Minimally, metamers that are generated for a natural speech stimulus should be **recognizable** to humans

Human recognition of model metamers



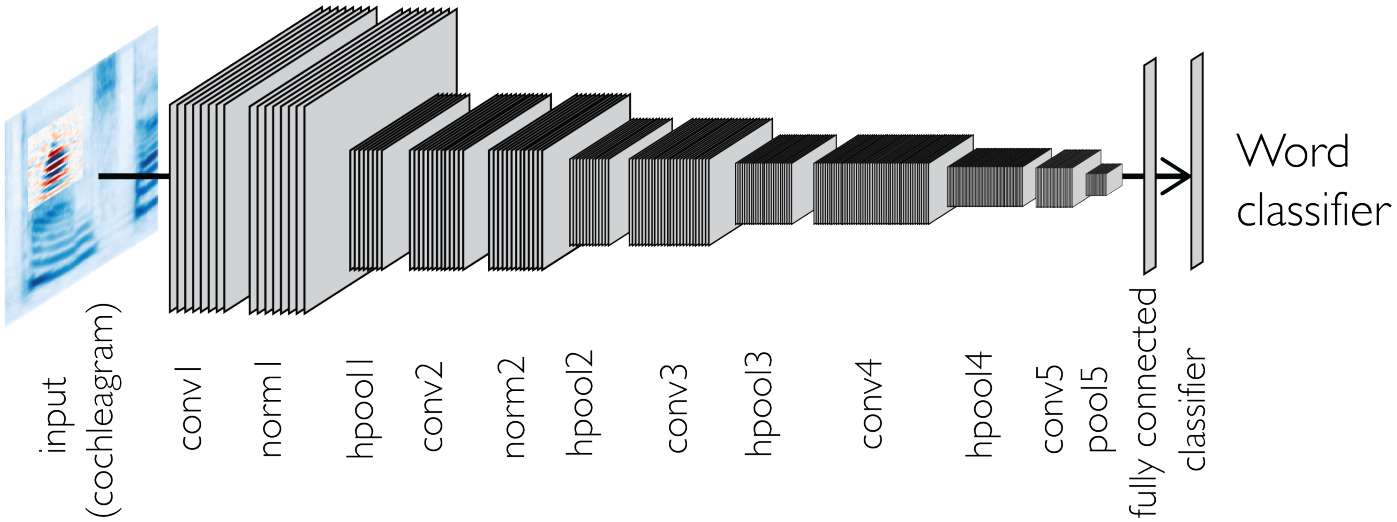
If humans are not able to recognize the model metamer the model invariances do not match human invariances.

If human responses are the same for original and model metamer, invariances may be shared between the two systems

Feather et al., NeurIPS (2019)

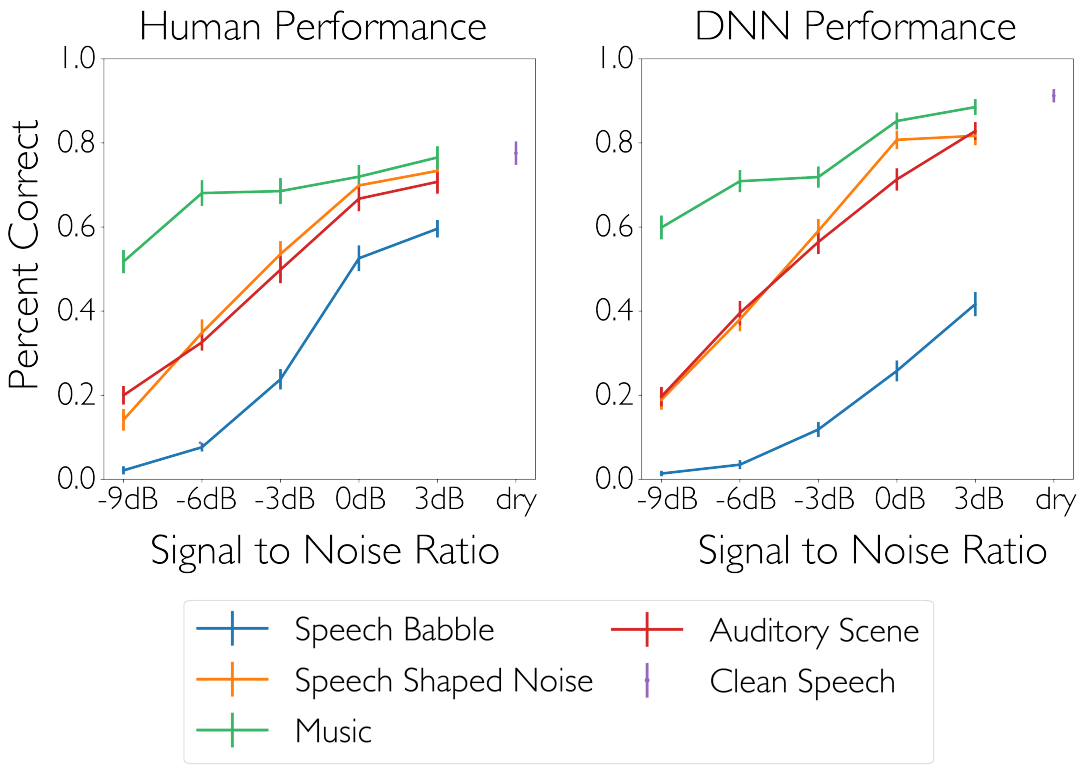
DNN models of human auditory system

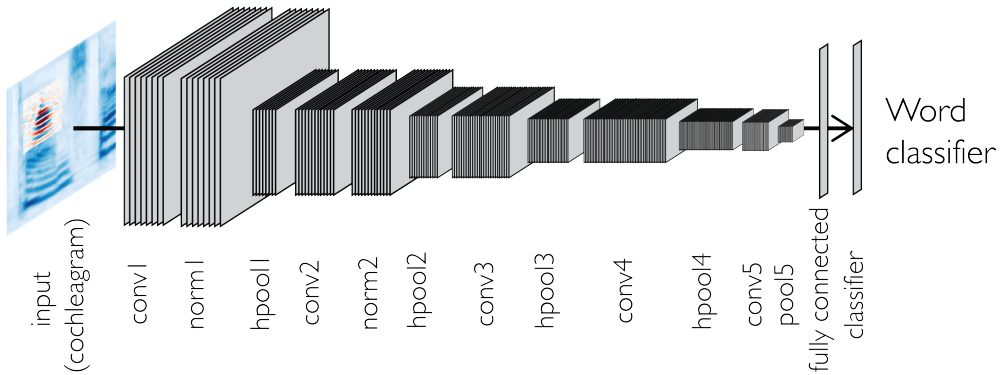
- Task: identify the word in the middle of a 2s sound clip containing background noise
- 793 possible words
- Natural sound background noise (AudioSet dataset)
- Input to network is a "cochleagram"



Architecture similar to Kell et al. 2018

DNN models of human auditory system



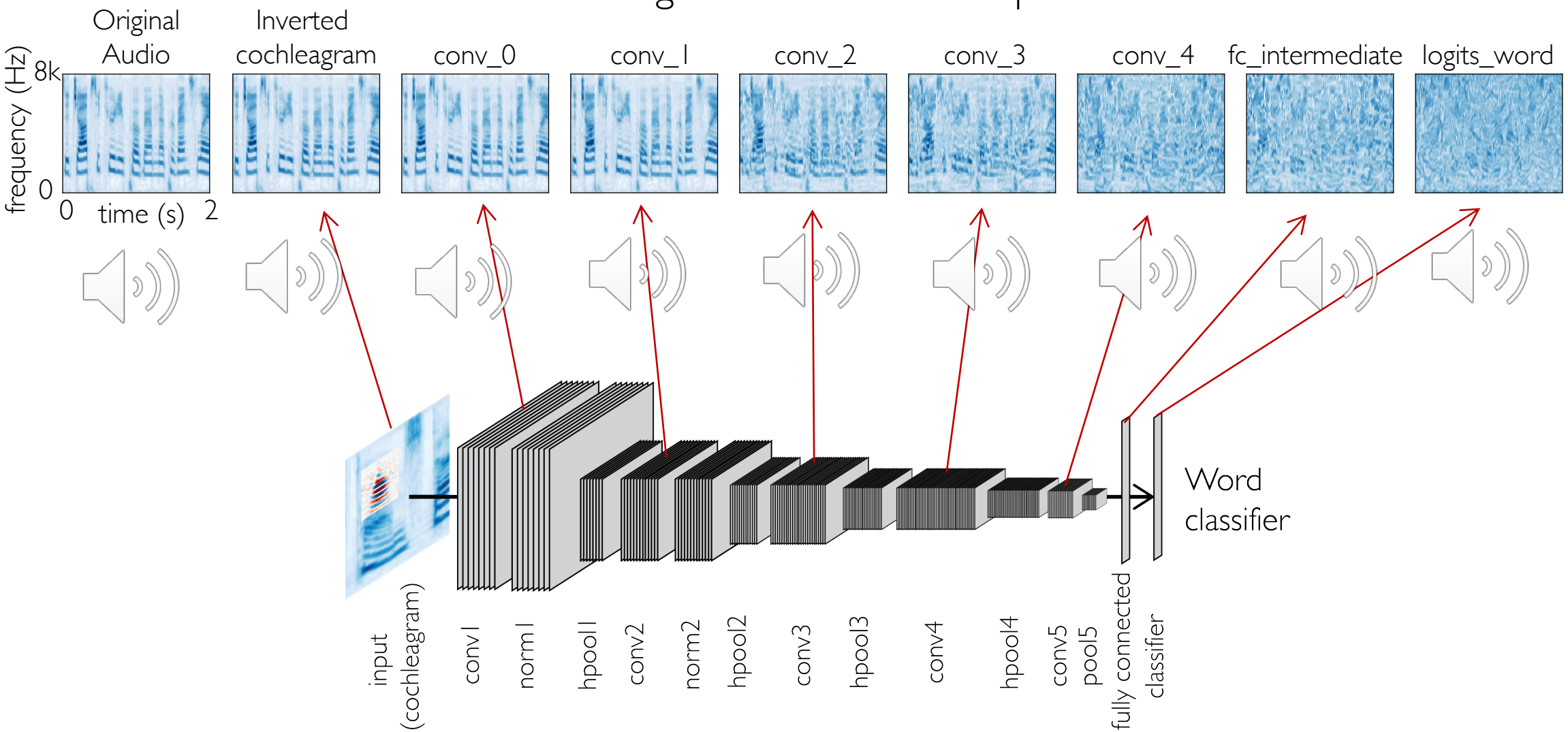


Key Question:
Do our DNN models of sensory systems share invariances with humans?

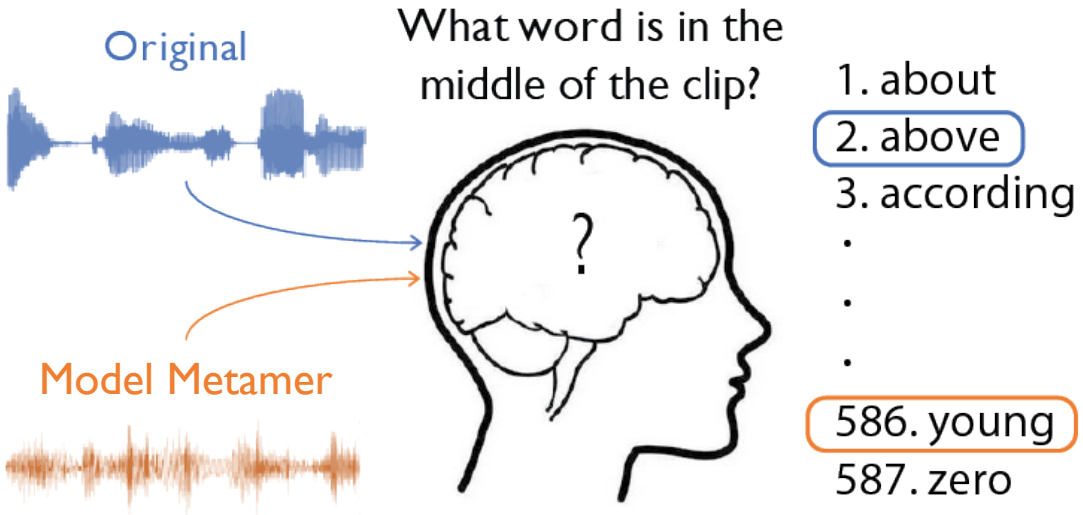
Architecture similar to Kell et al. 2018

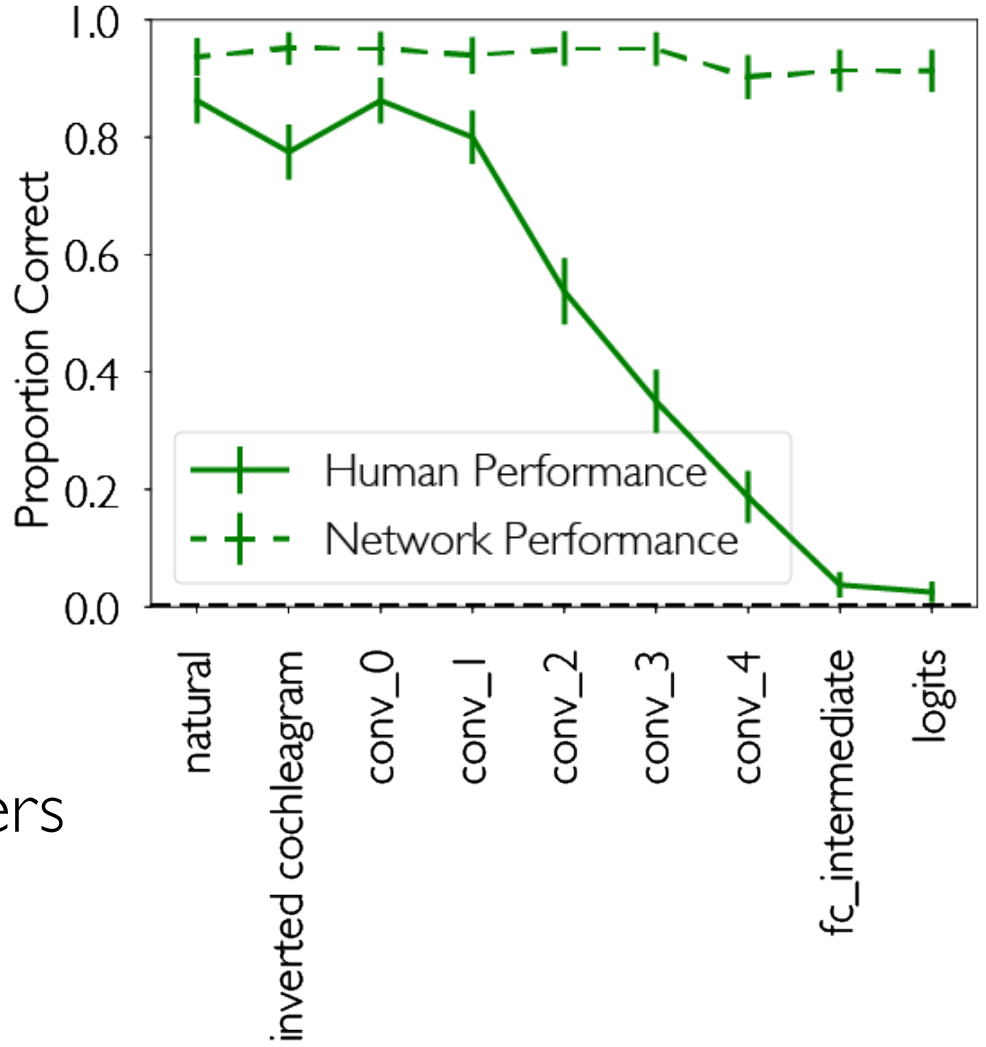
Human Behavior Data from Kell et al. 2018

Metamers generated for clean speech



Human behavior results **Audio Network**

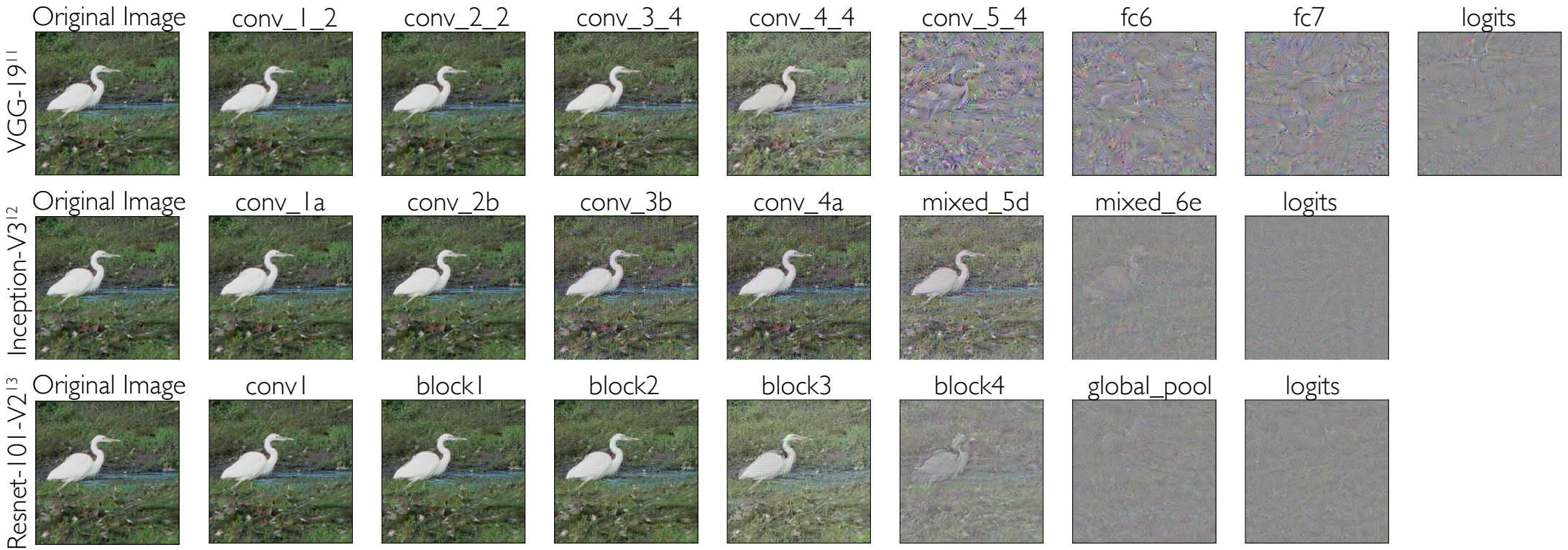




Model metamers from deep network layers are unrecognizable to humans

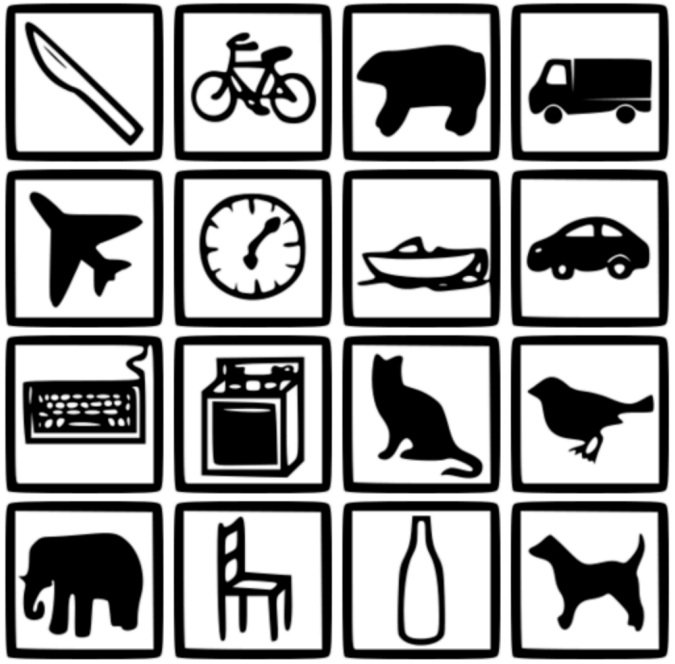
Feather et al., NeurIPS (2019)

Similar phenomenon for vision trained networks

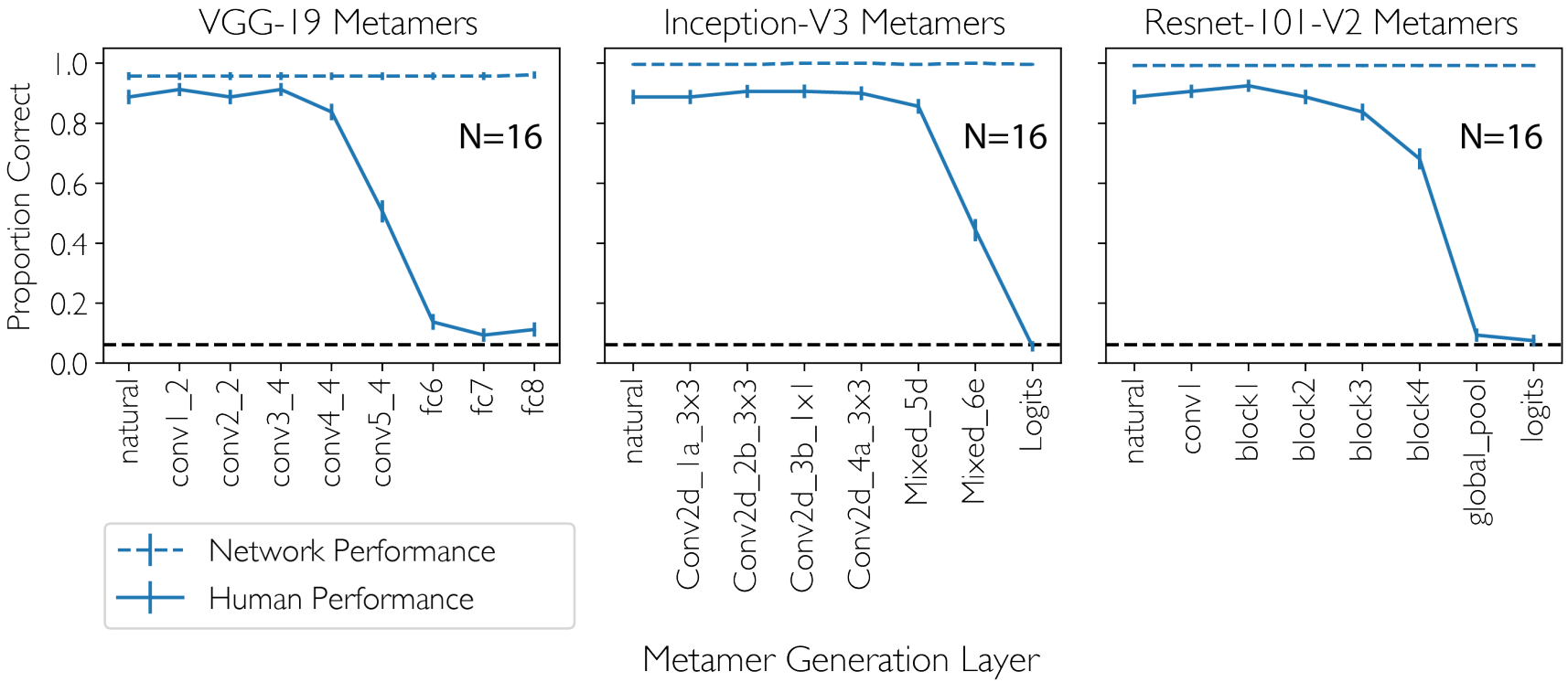


Feather et al., NeurIPS (2019)

Human behavior results **Image Networks**



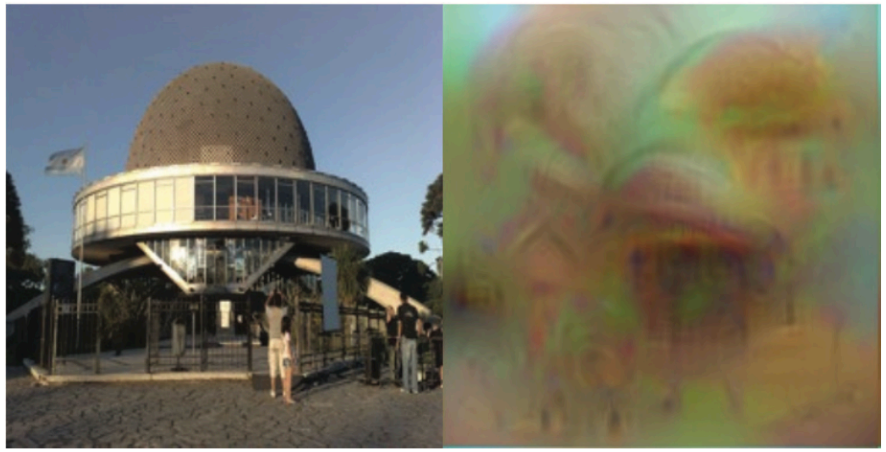
Task from Geirhos et al (2018)



Model metamers from deep network layers are unrecognizable to humans

Feather et al., NeurIPS (2019)

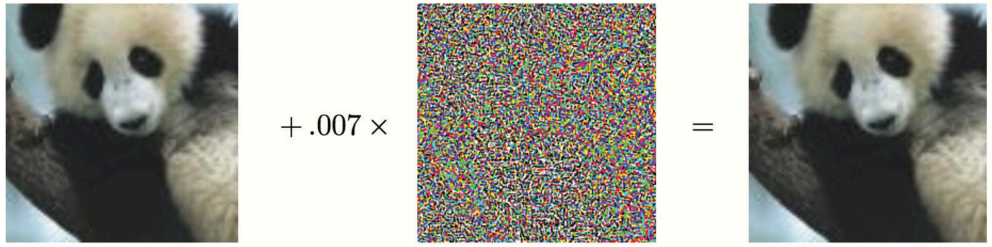
Mahenran & Vedalidi 2015



- Method of inverting the network representation is nothing new, but the link to perception has been under appreciated

- Most previous work relies on smoothness priors to make visually appealing images, which may hide model inadequacies

Goodfellow et al. 2015



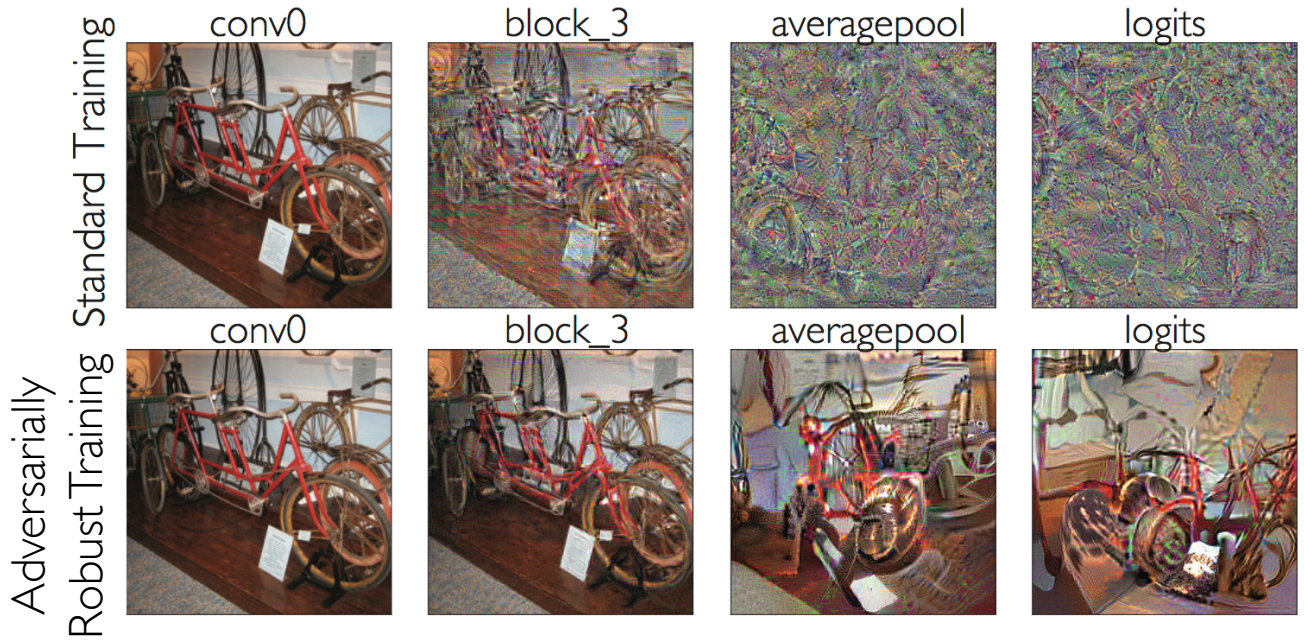
"Panda" "Gibbon"

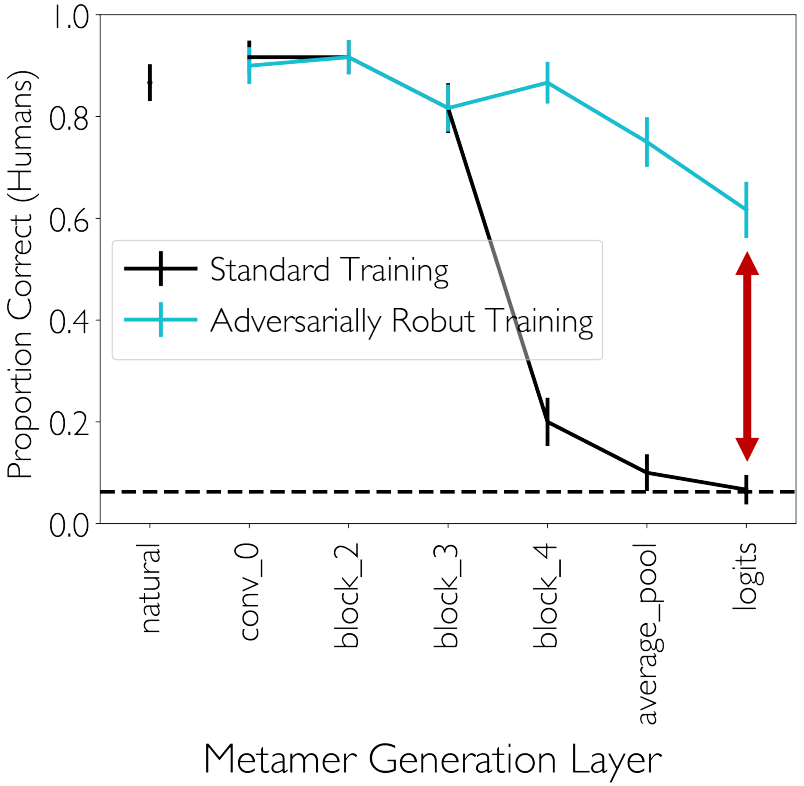
- Adversiarial examples are stimuli that are metameric for humans but are different for the network

Network invariances for auditory and visual DNNs do not match human perceptual invariances

Current Work:

How can we make our models better resemble human perception?

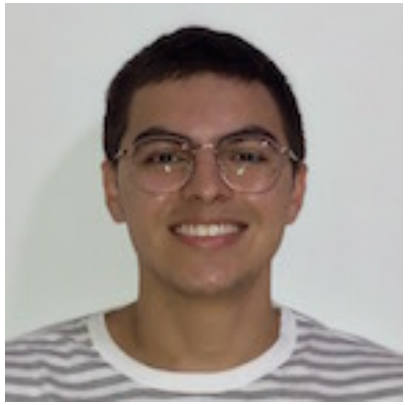




Collaboration with Guillaume Leclerc and Aleksander Mądry

Acknowledgements

Josh McDermott



Alex Durango

Ray Gonzalez

Adversarial Robustness Collaborators:

Guillaume Leclerc and Aleksander Mądry

Shared Code/Dataset:

Andrew Francl & Mark Saddler

Lab for Computational Audition

Funding Sources:

DOE CSGF Fellowship

McDonnell Scholar Award

NSF grant BCS-1634050

NIH grant R01-DC017970