

### Motivation

Deep neural networks (DNNs) have shown promise as models of sensory systems, for both vision and audition<sup>1,2</sup>

- Have human-like error patterns and can predict brain activity<sup>3,4</sup>
- Obvious failure cases by designing stimuli built to ''trick'' the model (adversarial examples)<sup>5.6.7</sup>
- •No obvious metric for how much the invariances of a model match human invariances

- a system
- are recognizable to humans or other models







**Network Visualization** 

Most previous work relies on smoothness priors to make visually appealing images, which may hide model inadequacies<sup>12</sup>



Visual Crowding/Texture perception average features to explicitly create human-like invariance<sup>8,9,10,11</sup>



Adversarial examples Metamers used to study human Metameric for humans but judged differently by the model (flip side of model metamers)<sup>5.6.7</sup>

# Metamers of neural networks reveal divergence from human perceptual systems

<sup>1</sup>Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology | <sup>2</sup>McGovern Institute, Massachusetts Institute of Technology <sup>3</sup>Center for Brains Minds and Machines, Massachusetts Institute of Technology | <sup>4</sup>Speech and Hearing Bioscience and Technology, Harvard University

#### Model metamers are not recognizable to humans

Audio: Human behavioral experiment<sup>3</sup>

Audio network training details:

- Word Task: Identify the word in the middle of the clip from 793 possible words<sup>3</sup>
- Audioset Task: Identify the environmental sounds in the clip from 516 sound categories<sup>13</sup>
- Training stimuli: Two seconds of speech superimposed on Audioset backgrounds

Metamer Generation Layer

• Model metamers matched to deeper layers of the ImageNet-trained networks are unrecognizable to humans

Main finding: Network invariances are not the same as human invariances

# Jenelle Feather<sup>1,2,3</sup> | Alex Durango<sup>1,2,3</sup> | Ray Gonzalez<sup>1,2,3</sup> | Josh McDermott<sup>1,2,3,4</sup>

# Additional audio model experiments

#### Metamers guide model improvements

- Network with reduced aliasing<sup>18,19</sup> (ensuring a low-pass filter before downsampling) has significantly more recognizable metamers
- Network architecture modifications can lead to internal representations closer to human perception



## Recognition of model metamers decreases after training



- Model metamers generated from a random network are more recognizable at late layers than those from a trained network
- Trained network, but not random network collapses noisy and clean speech to the same points



# Human recognition of model metamers is task dependent



- Network trained on Audioset task has model metamers of speech that are less recognizable to humans
- Although Audioset speech metamers are less recognizable, the sounds are less "noisy", suggesting further training changes could make models more human-like





### Model metamers as a model comparison tool

Comparison of audio networks trained on different tasks

• Model metamers can used to measure the similarity of internal representations of networks



- Training alters the network representations and different tasks lead to different invariances: metamers do not transfer between the Audioset and word trained networks
- Transfer between two random seeds provides proof of concept that metamers can be shared across distinct systems



#### Comparison of different image trained architectures



- Representations diverge for different ImageNet-trained architectures
- Point at which metamers become unrecognizable is similar to the fall off for humans

#### Discussion

- Model metamers are a tool for comparing computational models and biological systems
- Humans cannot recognize model metamers matched to late layers of a DNN, revealing a divergence between model representations and human perception despite similar behavior for natural stimuli in training set
- Model metamers reveal the invariances that are learned by a network, and provide an error signal to track when modifying models and training tasks

#### References and Acknowledgments

- [1] Yamins and DiCarlo (2016)
- [2] Kriegeskorte (2016)
- [3] Kell et al. (2018)
- [4] Rajalingham et al. (2015)
- [5] Goodfellow et al. (2014)
- [6] Berardino et al. (2019)
- [7] Jacobsen et al. (2019)
- [9] McDermott and Simoncelli (2011) [16] Szegedy et al. (2016)
- [10] Portilla and Simoncelli (2001) [17] He et al. (2016)
- [1] Balas et al. (2009)
- [12] Mahendran and Vedaldi (2015) [19] Azulay and Weiss (2019)
- [13] Gemmeke et al. (2017)
- [14] Geirhos et al. (2018)
- [8] Freeman and Simoncelli (2011) [15] Simonyan and Zisserman (2014)

  - [18] Zang (2019)

Funding Sources: McDonnell Scholar Award to J.H.M. and NSF grant BCS-1634050 to J.H.M. DOE CSGF Fellowship to J.J.F