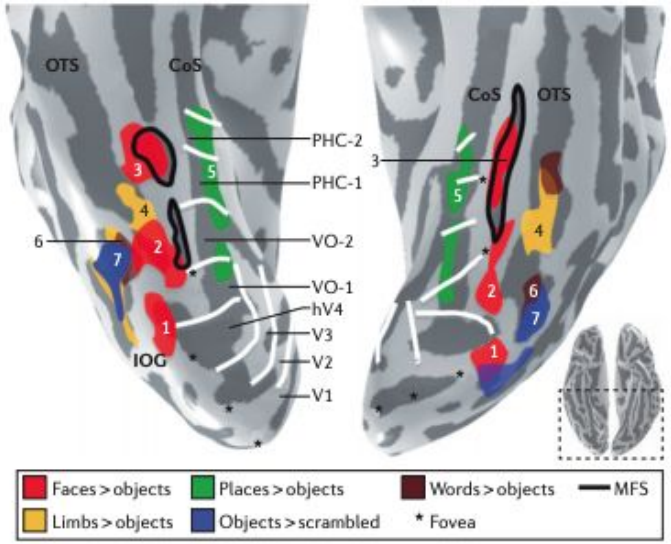
Does a DCNN have category-selective representations?

Violet Xiang & Emily Kubota

Background (E)

- In humans lesions in ventral temporal cortex can result in deficits in object recognition.
- There is debate about whether these lesions are category-specific or affect multiple categories.
- Human patients vary in the size and spatial location of lesions within VTC.
- Obviously, it's unethical to lesion patients systematically.



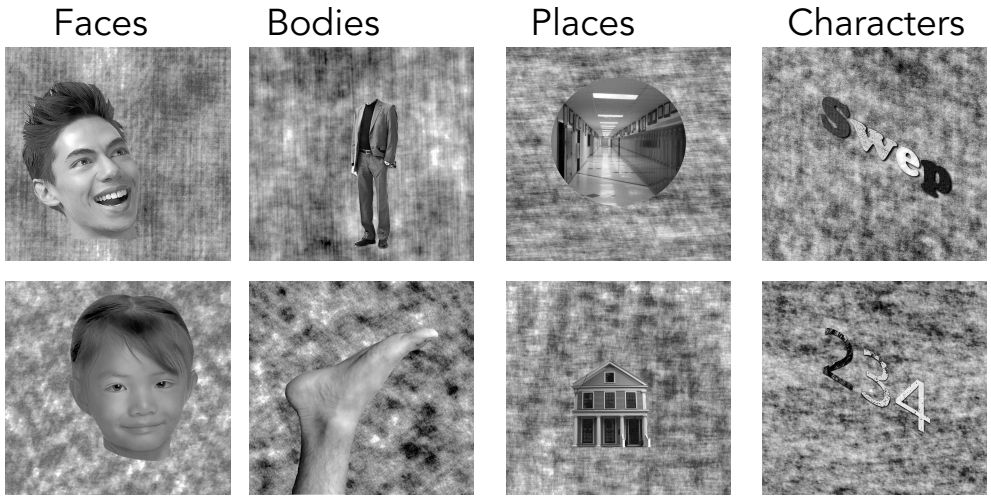
Weiner & Grill-Spector, 2014

Current project (E)

- Goal: lesion a DNN and test how it affects image classification.
- Research questions:
  - Will lesioning a network selectively affect classification of some categories and not others?
    - Similar to a “double-dissociation” in patient studies
  - Will lesioning different layers differentially affect classification of different categories?
    - Suggesting category-selective regions
  - Will lesioning different epochs affect classification of categories differently?
    - Suggesting different onset of category-selective representations in development

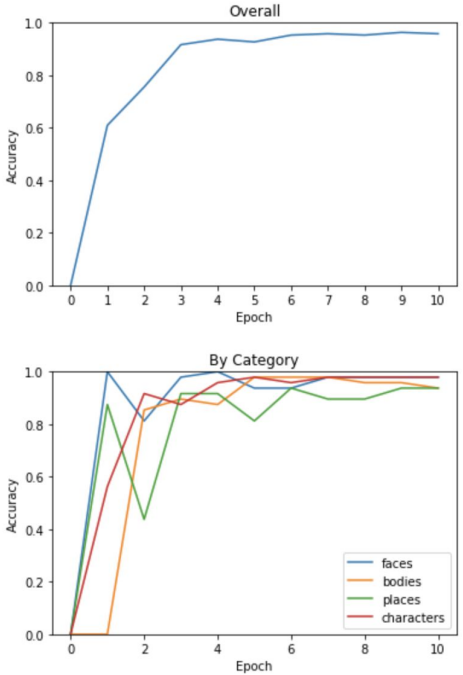
Experiments (V)

- Model: AlexNet pre-trained using ImageNet
- Lesioning the network
  - 20% filters in each convolution layer are randomly selected as lesion targets for 10 times
  - Evaluation of overall performance and individual category performance

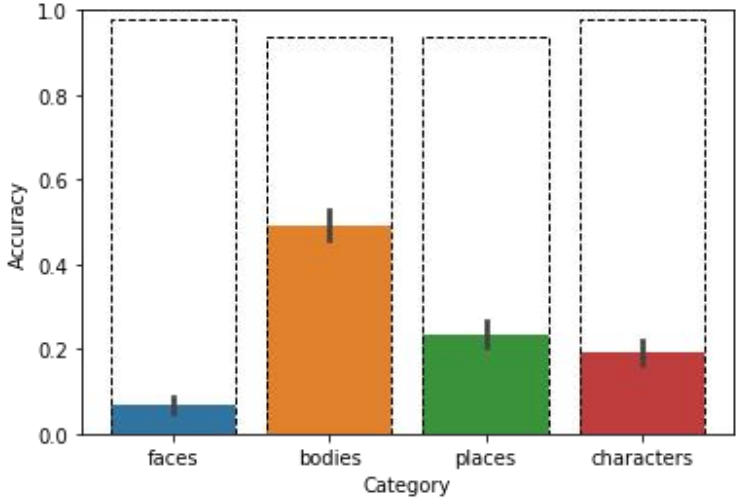


Stigliani et al., (2015)

Network without lesion performs well on the task (E)



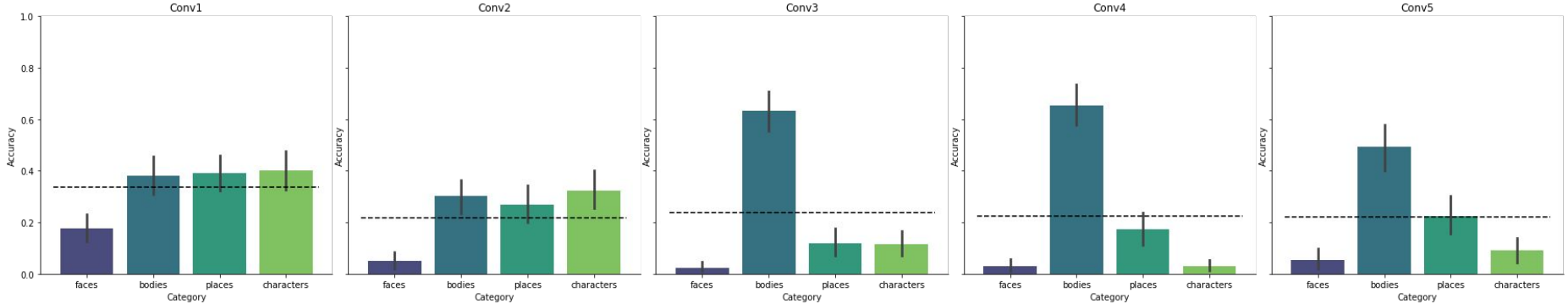
Does lesioning the network affect some categories more than others? (E)



Face classification was impacted the most and body classification was impacted the least.

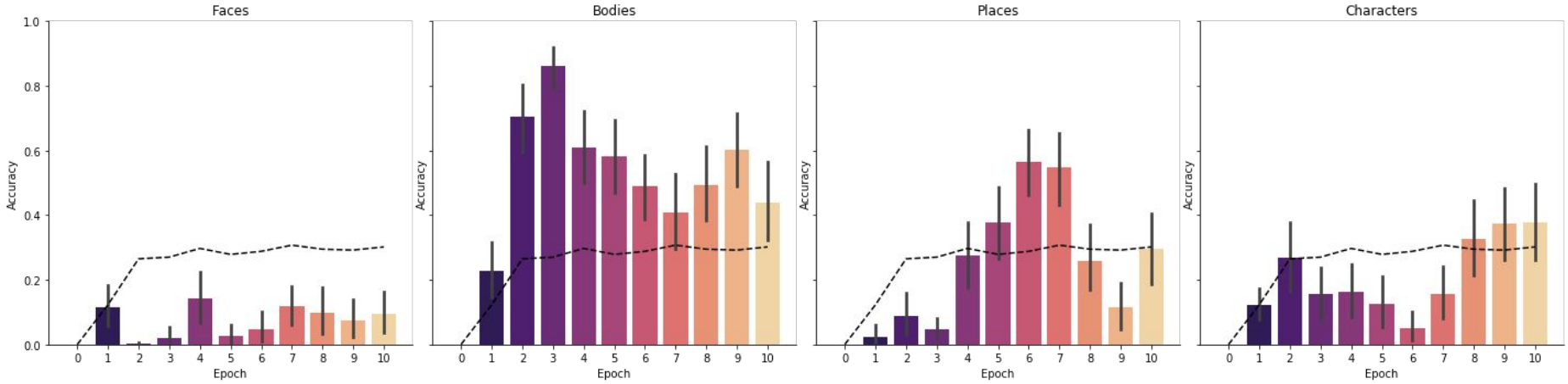
= accuracy on the network without lesions

Layers that affect classification the most are different for different categories (V)



------- = mean accuracy for each layer with lesions
(not separated by category)

Do category-selective representations emerge early or late in training? (V)



------- = mean accuracy for lesioned network at each epoch (not separated by category)

Conclusion (V)

- Lesioning has different effects on different categories
- Different category representations are affected at some layers more than others, suggesting that category-specific representations are distributed different across layers
- Limitations:
    - DNNs are not topographical like the human visual system