Using Human Brain Activity to guide Machine Learning Fong et al.

ABNS 2023

Outline

- 1. Introduction + context
  - a. Intro to methods
  - Intro to results
- 2. Methods + results
  - fMRI activity weight calculation
  - Experimental findings
- 3. Discussion + issues
  - Related work
  - Limitations

Introduction

Brain areas, Features, and Loss functions.

Guiding questions and premises

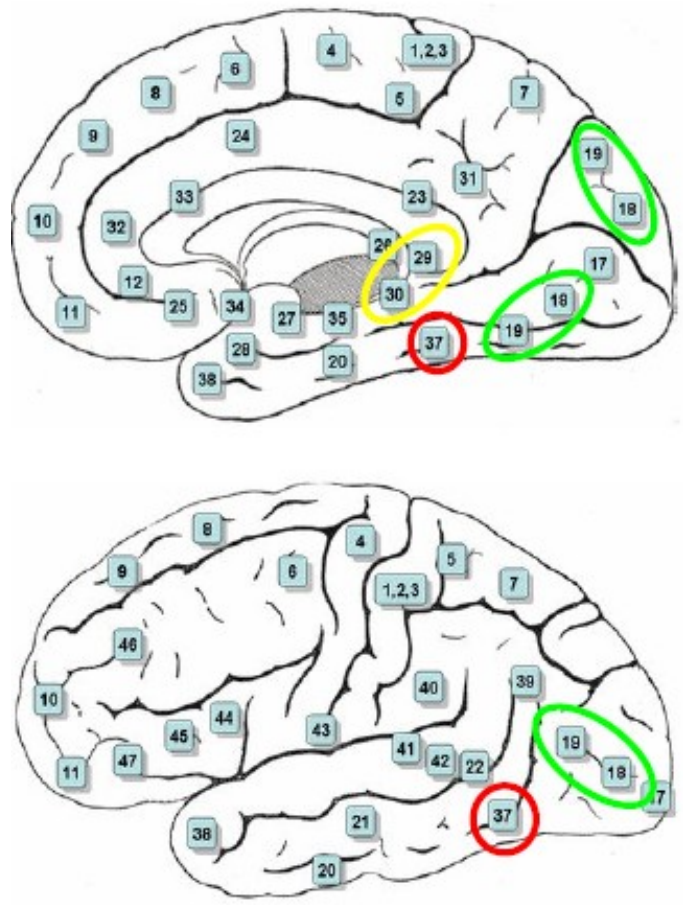
- Is it possible to improve ML performance by guiding performance with brain activity
- Will such guidance make the representations themselves more ‘human like’?
  - Here, the focus (would be) not on performance but on representational geometry
- If the human brain is a natural reference point (**for representation geometry) and performance (**that counts), and ANNs are a good algorithm for learning structure, we can attempt to leverage the ML algorithm with biological information.

Brain areas considered

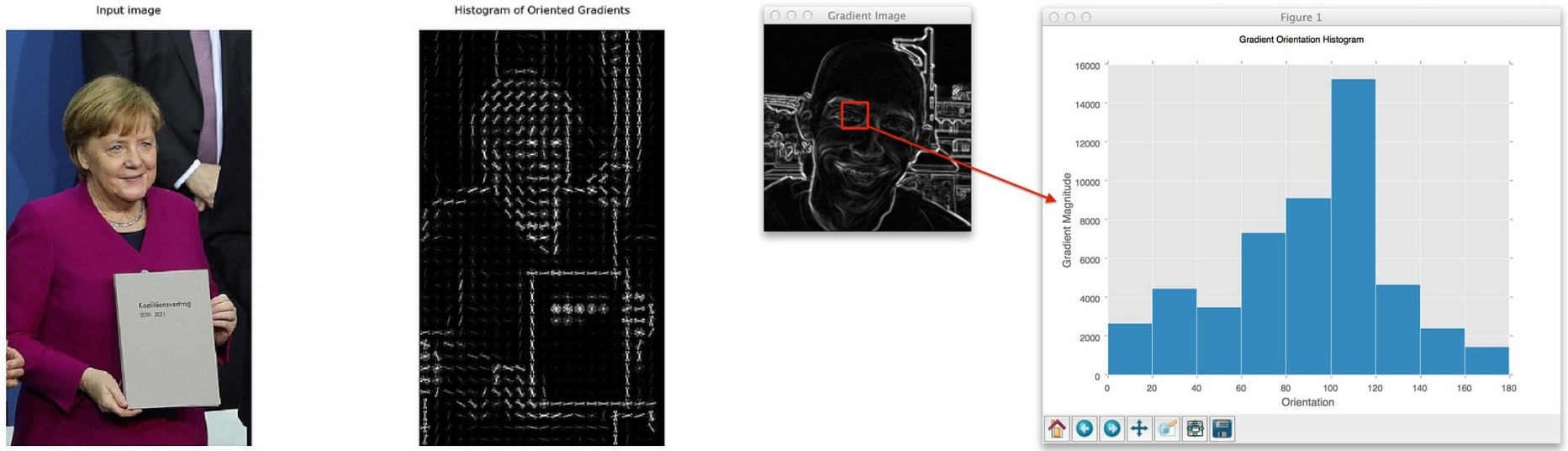
- Different brain areas code for different information about the stimulus.
- Here, the partitioning of brain areas that will supervise ML is done a-priori.
- They focus mainly on visual regions

7 ROIs associated with high-level visual understanding

- EBA, extrastriate body area
- FFA, fusiform face area
- LO, lateral occipital cortex
- OFA, occipital face area
- PPA, parahippocampal place area
- RSC, retrosplenial cortex
- TOS, transverse occipital sulcus


4
1,2,3
6
8
5
7
9
24
19
31
33
23
18
10
32
29
17
12
30
18
11
25
34
27
35
19
37
28
20
38
8
4
5
9
6
1,2,3
7
46
40
39
10
44
43
45
41
42
22
19
18
11
47
17
37
21
38
20

Types of models used: HOG and CNN



- The histogram of oriented gradients (HOG) is an often-used algorithm for generating features that capture the local “slopeness” of different parts of an image
- Convolutional NN.

Hinge Loss

Typically implemented in binary classifiers. Aim: Maximize the margin between the decision boundary and the data points.

Denote ‘t’ as true label; t = 1 or -1
Denote ‘y’ as predicted activation for object by classifier
t * y is the distance from the boundary. The Hinge loss wants to push this distance so that the margin from the boundary is above 1.

*Loss(y) = max(0, 1 – t*y)*

Decisions that incur loss (incorrect classification):

- *Prediction is on incorrect side of boundary and so, t and y mismatch in signs.*
  - *(t * y) < 0*
  - *1 – t*y will always be positive and above 1*
- *Prediction is on correct side of boundary (t and y match in signs) by y < 1.*
  - *0 < (t* y) < 1*
  - *(1 – t*y) > 0, and so reflects a loss*

Decisions that do no incur loss (“correct”)

** Ones where |y| is greater than the margin (1): If (t = 1 & **y >1**) or (t = -1 & **y < -1**) , **t*y > 1** and so: 1 – t*y **< 0***

Loss is proportional to the distance from the SVM’s decision margin.

Important to remember: the ‘hinge’ (max) ignores the distance of correct decisions (points) from the boundary (distance > 1).

What matters are the closest points in the two classes and how they should be separated.

They separation plane minimizes errors, does not care about magnitude of correct decisions **above the margin** (TBD)

Activity Weighted Loss

- Prelim: defined ‘response strength’ From brain fMRI activity data as the distance of an object from the decision boundary for a given binary classification task.
- This produces a per-stimulus activity weight (response strength) for each stimulus.
- As input to the classifier, each image is encoded as a vector of brain-activity values sampled from a given brain area.
- No need to know the details of how the fMRI data is processed.

Hinge Loss (HL)

$$\phi(x, z) = max(0, 1 - z)$$

$$\text{with } z = yf(x)$$

(t * y)

Penalizes misclassified examples.

Activation Weighted Loss (AWL)

$$\phi(x, z) = max(0, (1 - z) * M(x, z))$$

$$M(x, z) = \begin{cases} (1+C_x) \text{ if } z<1 \\ 1 \text{ otherwise} \end{cases}$$

Penalizes misclassified examples on stimuli that are easy for humans to distinguish.

Note: 'X' here is the classification prediction; further penalization based on human knowledge **only applies** when (Z < 1) [(t * y) < 1]

Hinge Loss (HL)

$$\phi(x, z) = max(0, 1 - z)$$

with $z = yf(x)$

Penalizes misclassified examples.

Activation Weighted Loss (AWL)

$$\phi(x, z) = max(0, (1 - z) * M(x, z))$$

$$M(x, z) = \begin{cases} (1+C_x) \text{ if } z<1 \\ 1 \text{ otherwise} \end{cases}$$

Penalizes misclassified examples on stimuli that are easy for humans to distinguish.

- What is the impact on ‘representational geometry’ or class representation?
- What changes? In/out members; centroid. Identity of boundary members.

TBD

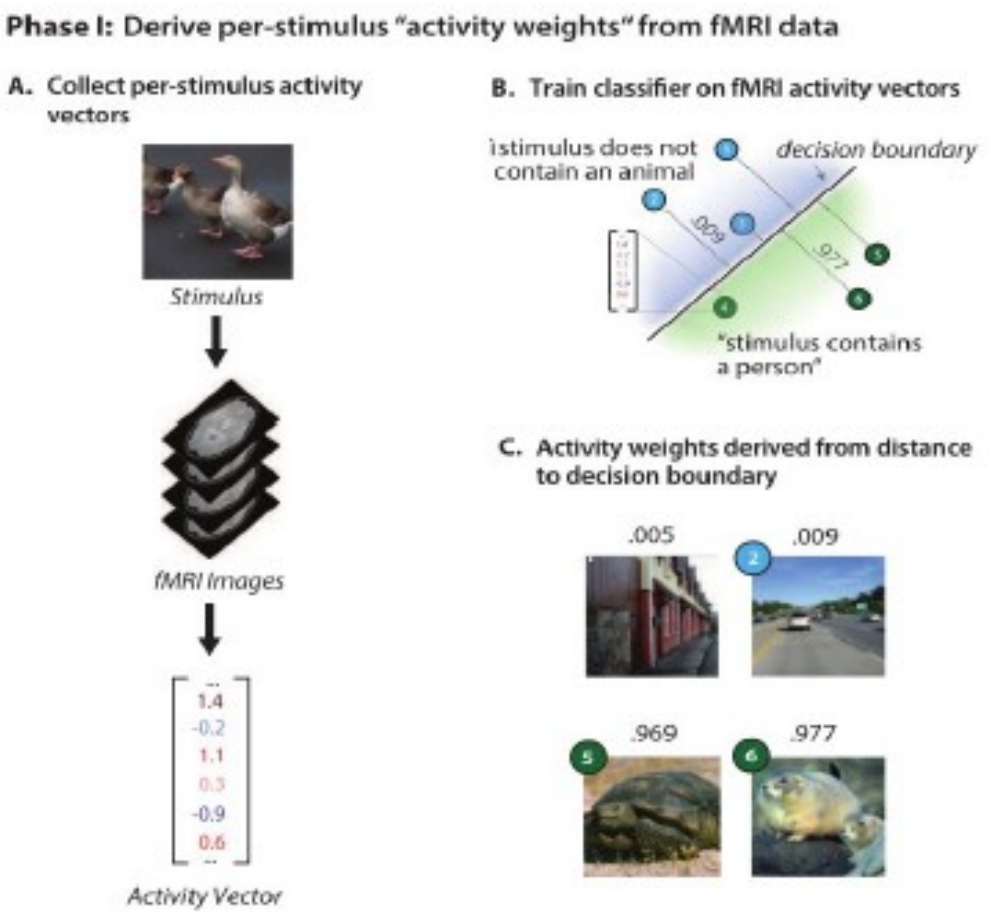
AWL: Step 1.

Experimental workflow - 1

A. *Collect per-stimulus activity vectors*: use fMRI to record BOLD response of subject;

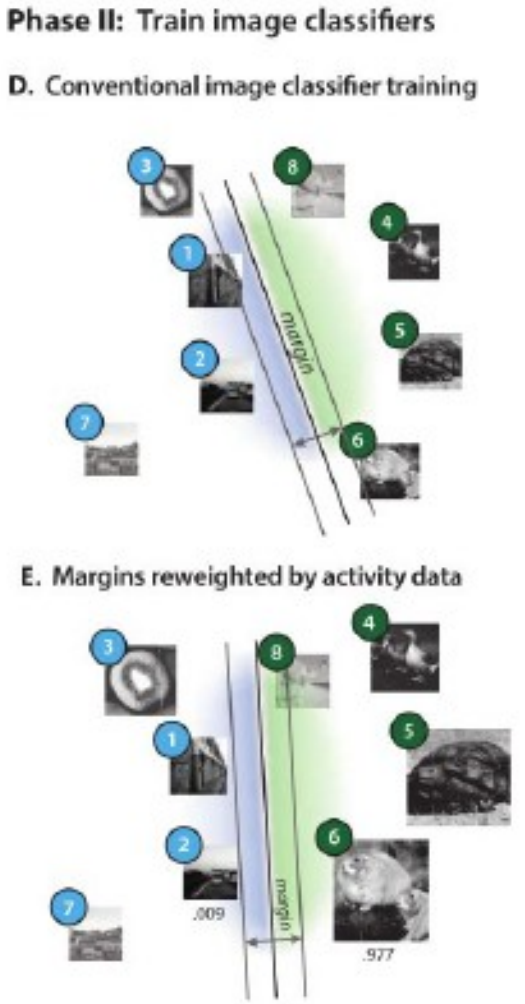
B. *Train classifier on fMRI activity vectors*: SVM classifier trained and tested;

C. *Activity weights derived from distance to decision boundary*: use transformed classification scores.



AWL: Step 2.

- *D: Conventional image classifier training*: Radial Basis Function SVM classifier
- *E: Margins reweighted by activity data*: SVM classifier with activity weighted loss function. NOTE: not all training samples require fMRI weight.

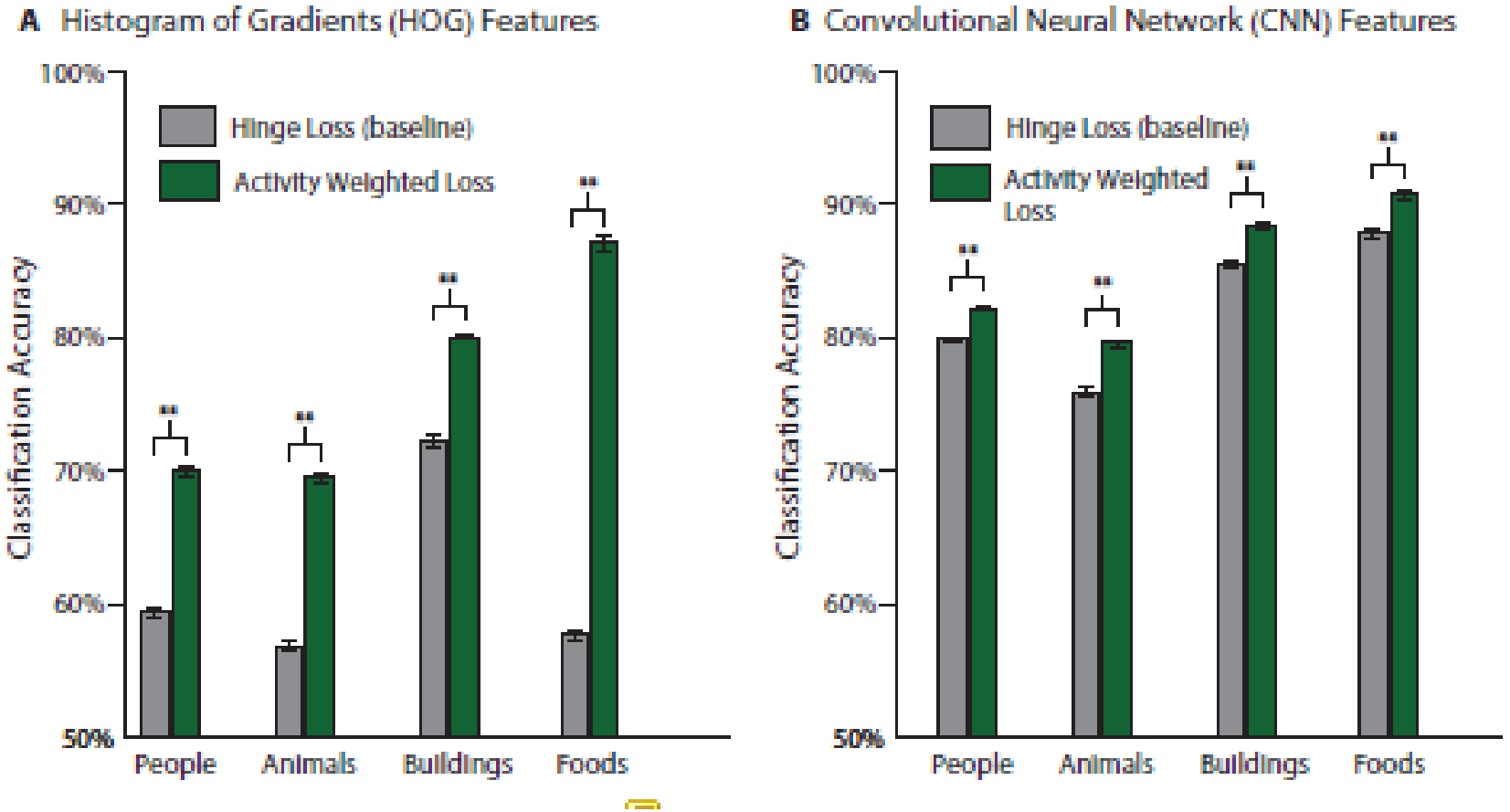


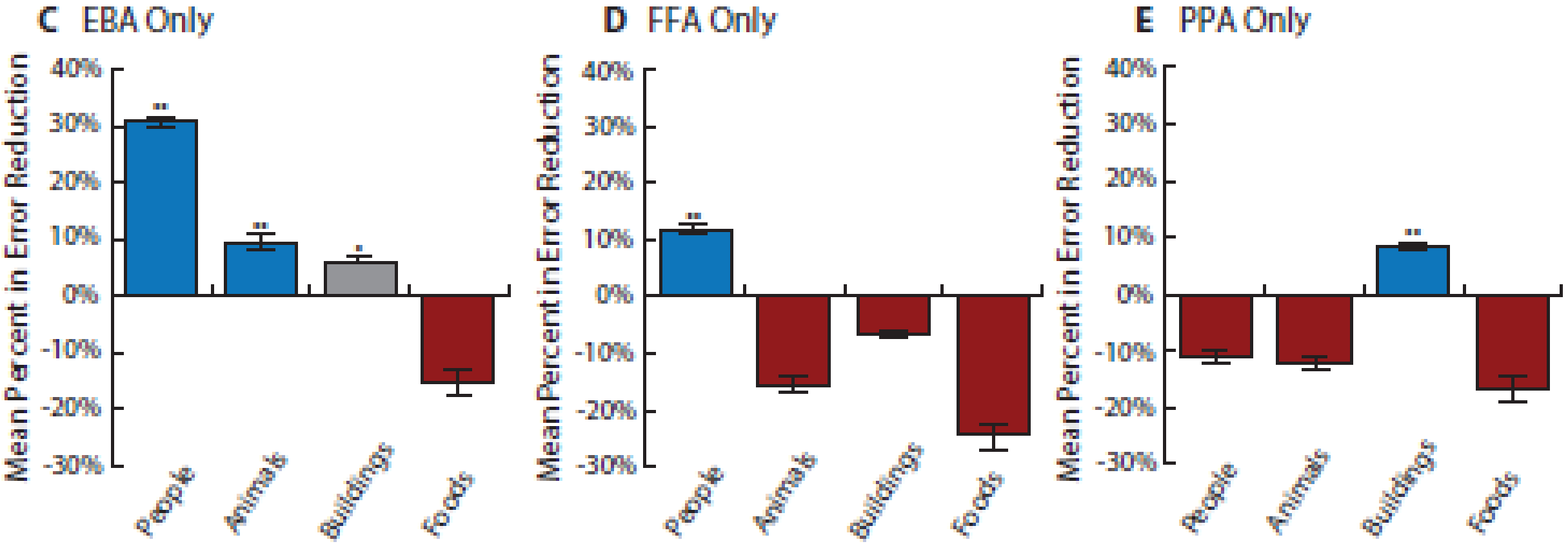
Other information

- Use images from 5 categories
  - Humans -> 219 images
  - Animals -> 180 images
  - Buildings -> 151 images
  - Foods -> 59 images
  - Vehicles -> 37 images
- Classification problems based on CNN or HOG features.
- Information from the higher-level cortical regions combined in all possible combinations to produce feature sets.

Results

Classification accuracy: all brain areas used together





Does information from different brain areas differentially discriminate different information?

- EBA, FFA, and PPA == body parts, faces, places?
- “Given the overlap between these visual cues and the four object categories used, we hypothesized that activity weights derived from brain activity in these three regions would significantly improve classification accuracy for the humans, animals, and buildings categories”

Conclusions

- Information measured directly from brain can guide an ML algorithm “to make better human-like decisions”.
- One can harness measures of the internal representations employed by the brain to guide machine learning.
- Question: is the importance brain data, or more data? What if we improved the HOG classifier using AWL from CNN classifier?