Evaluating Reconstruction Accuracy in Neural Decoding Models

Team: Motorized Marmots
Members: Harmandeep Singh Khera & Boyin Feng
Pod: Gazpacho
Mentors: Ahmad Samara, Amy Kuceyeski, Javier Guaje

Introduction

- Decoding models relate regional brain activity (e.g., voxel-wise BOLD response) to the stimulus:

- In human vision research, decoding models are evaluated for their ability to use fMRI responses to an image for predicting:
  - its stimulus class (‘classification’)
  - recognise it among distractors (‘identification’),
  - or reconstruct the original image

- However….

  The “one pixel attack”

From Kay, K., Gallant, J. Nat Neurosci (2009)

Introduction (contd.)

- There are multiple metrics for quantifying model performance on image reconstruction — leading to:
  - ad-hoc choices
  - less comparability among decoding models

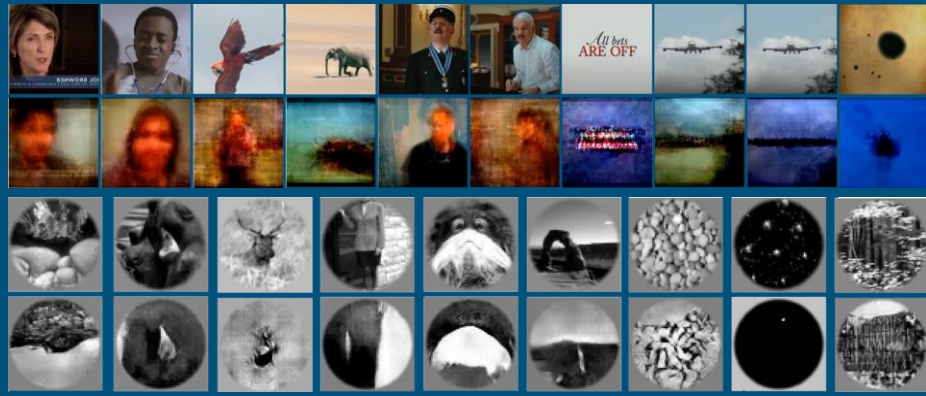
Study	Metrics Used
Bashivan et al., 2019	- Distance in pixel-space - Euclidean distance
Miyawaki et al., 2008	- Spatial correlation - Image identification
Naselaris et al., 2009	- Spatial similarity (Brooks and Pappas, 2006) - Semantic similarity, at 4 different levels of specificity

- We propose that human judgement should serve as the golden standard; we, therefore, compared 2 metrics of reconstruction accuracy against 2 human judgement tasks

Methods

- Human judgment:
  - Participants (N = 6, 3F, 3M; Age range: 23-27)
  - 2 Dimensions: accuracy; perceived similarity

- Conventional measures:
  - Metrics: RMSE (Root Mean Squared Error) and FSIM (Feature Similarity Index)
  - Implemented in Python using the image_similarity_metrics package (Müller et al., 2020)



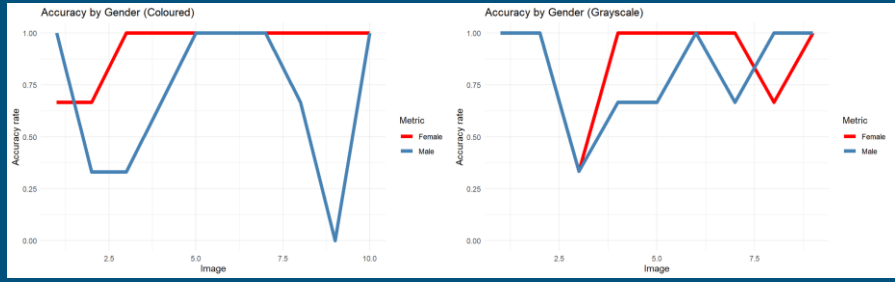
Materials:

- 10 colored pairs (Nishimoto et al., 2011) in 3 min;
- 9 grayscale pairs (K. Seeliger et al., 2018) in 2 min

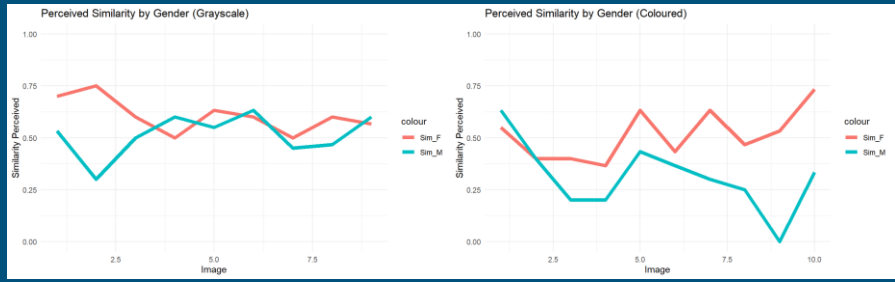
Results

Interesting results

- In general, females matched the reconstructed images with the original ones more accurately and more quickly
- Males perform better with grayscale images
- Mean similarity ratings were lower for males than females



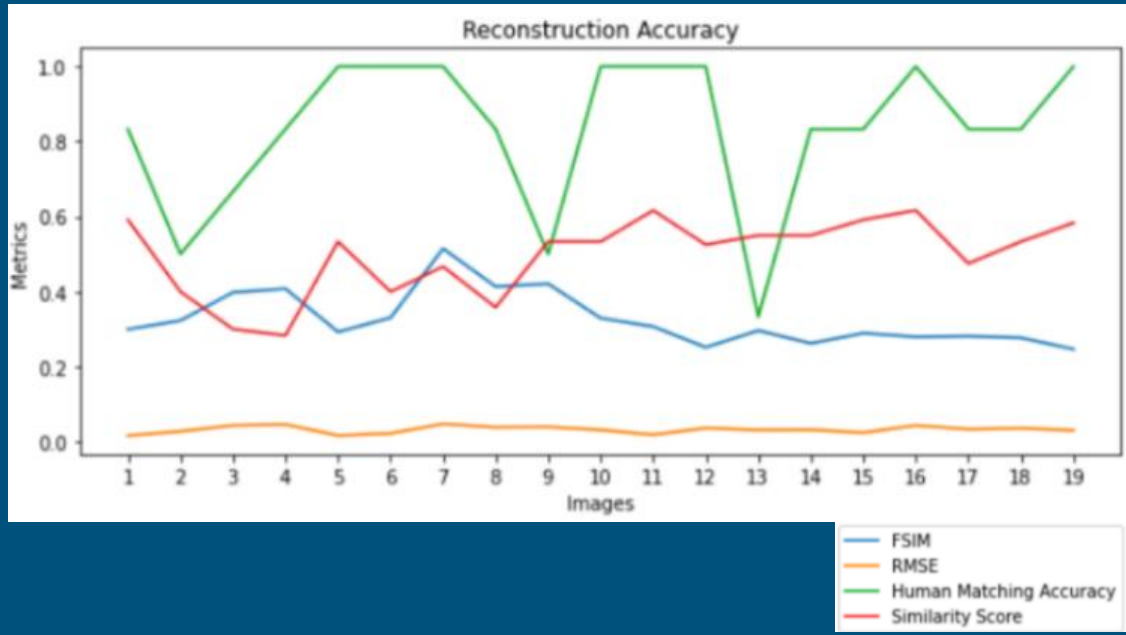
Accuracy by Gender



Perceived Similarity by Gender
(correct trials only)

Discussion

- Our results suggest that human judgements may help improve comparability among decoding models
- Future empirical surveys should investigate:
  - Differences b/w category of objects (e.g. faces vs tools)
  - Graded judgements of similarity (e.g. shape vs taxonomic class)
  - The metrics for human judgment need further examination as well



	Matching Accuracy	Similarity Score
RMSE	ρ = -0.12, p > 0.63	ρ = -0.44*, p > 0.055
FSIM	ρ = -0.14, p > 0.55	ρ = -0.58**, p < 0.01

***Almost** significant at p < 0.05

**Significant at p < 0.01