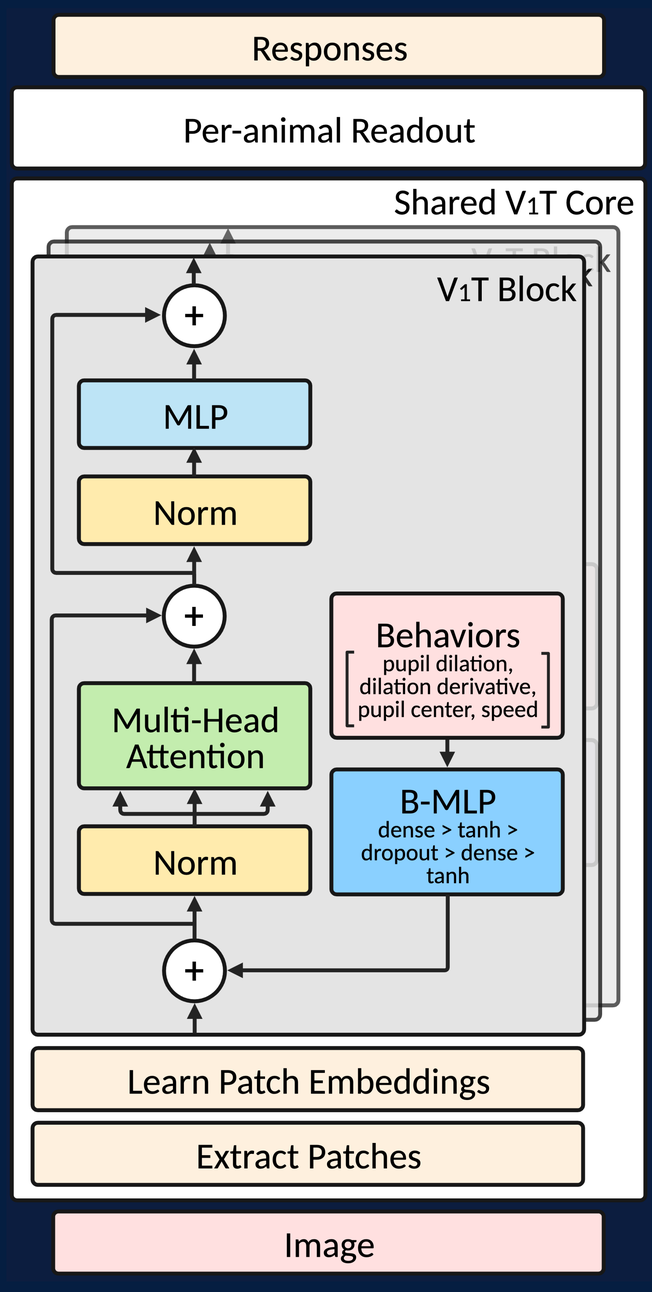
TL;DR: We propose a Transformer-based architecture that can accurately predict mouse V1 responses to natural images. We showed that the model learns notably narrower aRFs than its CNN counterpart, and the self-attention weights correlate with pupil directions.



V1T: large-scale mouse V1 response prediction using a Vision Transformer

Bryan M. Li[1], Isabel M. Cornacchia[1], Nathalie L. Rochefort[2,3], Arno Onken[1]

[1]School of Informatics, University of Edinburgh
[2]Centre for Discovery Brain Sciences, University of Edinburgh
[3]Simons Initiative for the Developing Brain, University of Edinburgh

Introduction

- Predictive models of neural responses to natural stimuli serve the dual purpose of generating new hypotheses about biological vision and bridging the gap between biological and computer vision [1].
- Data-driven approaches are dominated by CNN-based models though few studies investigate how to integrate behavioral information [2, 3].
- As Vision Transformers (ViT) [4] are becoming increasingly popular in computer vision, can an even less biologically plausible model be good at visual response prediction?

Task

- Given visual stimuli (natural images) and behavioral information, predict calcium responses from ~8k V1 neurons per animal (5 and 10 mice in [5] and [6]).

Method

- We proposed V1T, a ViT-based core architecture that learns a joint visual and behavioral representation across animals using a block-wise behavioral integration
- Compare against the previous SOTA CNN [2] on the two large-scale mouse V1 datasets under the same condition.

Prediction performance

Single trial correlation (averaged over 5 and 10 rodents, SD shows the standard deviation) in the two test sets. Additional result in cross-animal/dataset generalization, sample efficency and model ensemble are available in the paper!

	Sensorium2022 [5]		Franke et al. 2022 [6]	
Model	Trial Corr (SD)	ΔCNN	Trial Corr (SD)	ΔCNN
LN	0.275 (0.019)	-27.2%	0.223 (0.040)	-28.0%
CNN	0.378 (0.021)	0%	0.309 (0.070)	0%
ViT	0.414 (0.032)	9.5%	0.344 (0.041)	11.4%
V1T	**0.426 (0.027)**	**12.7%**	**0.368 (0.032)**	**19.1%**

Discussion

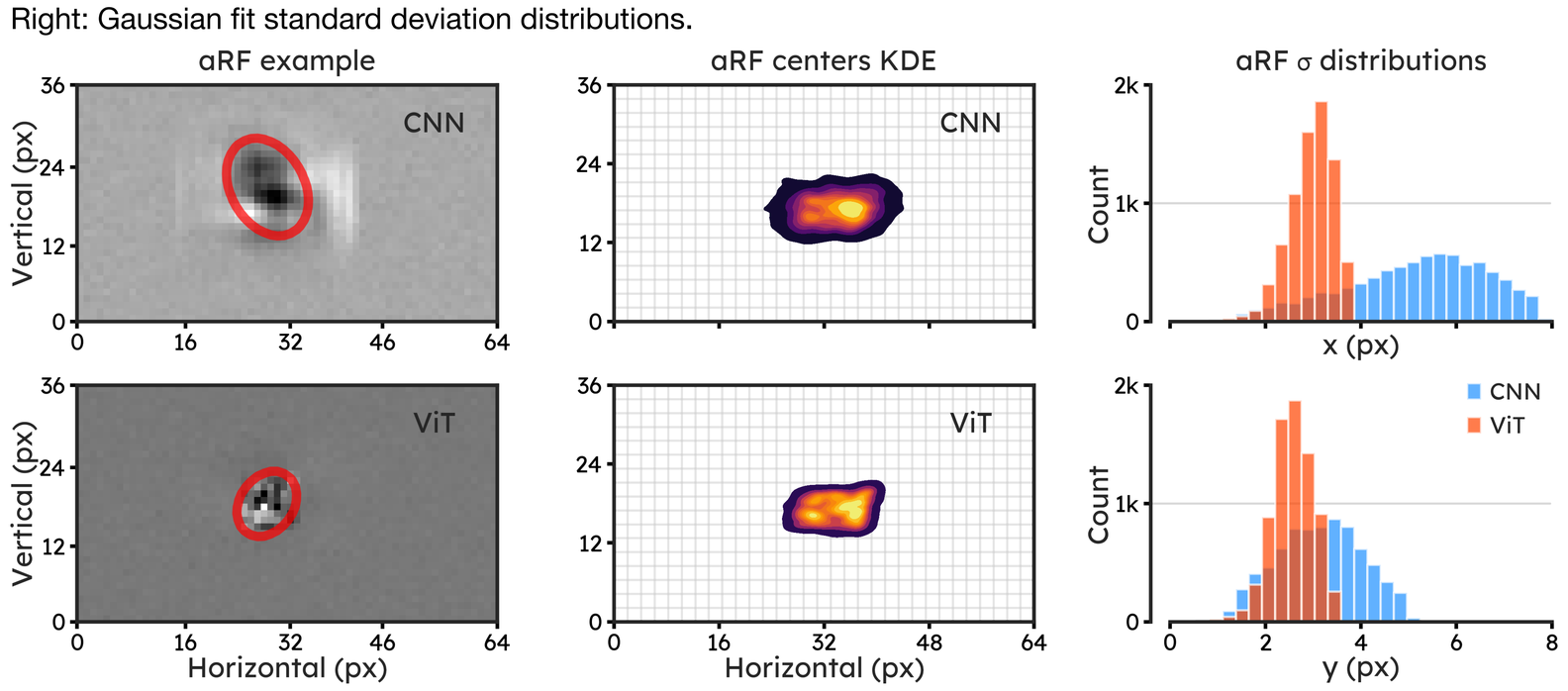
- The first ViT-based model to outperform CNN in mouse V1 prediction and provides a framework to investigate in silico computations in the visual system.

Future Work

- Investigate the relationship between behavioral variables and neural response (e.g. ablation).
- Extend model to dynamical settings.

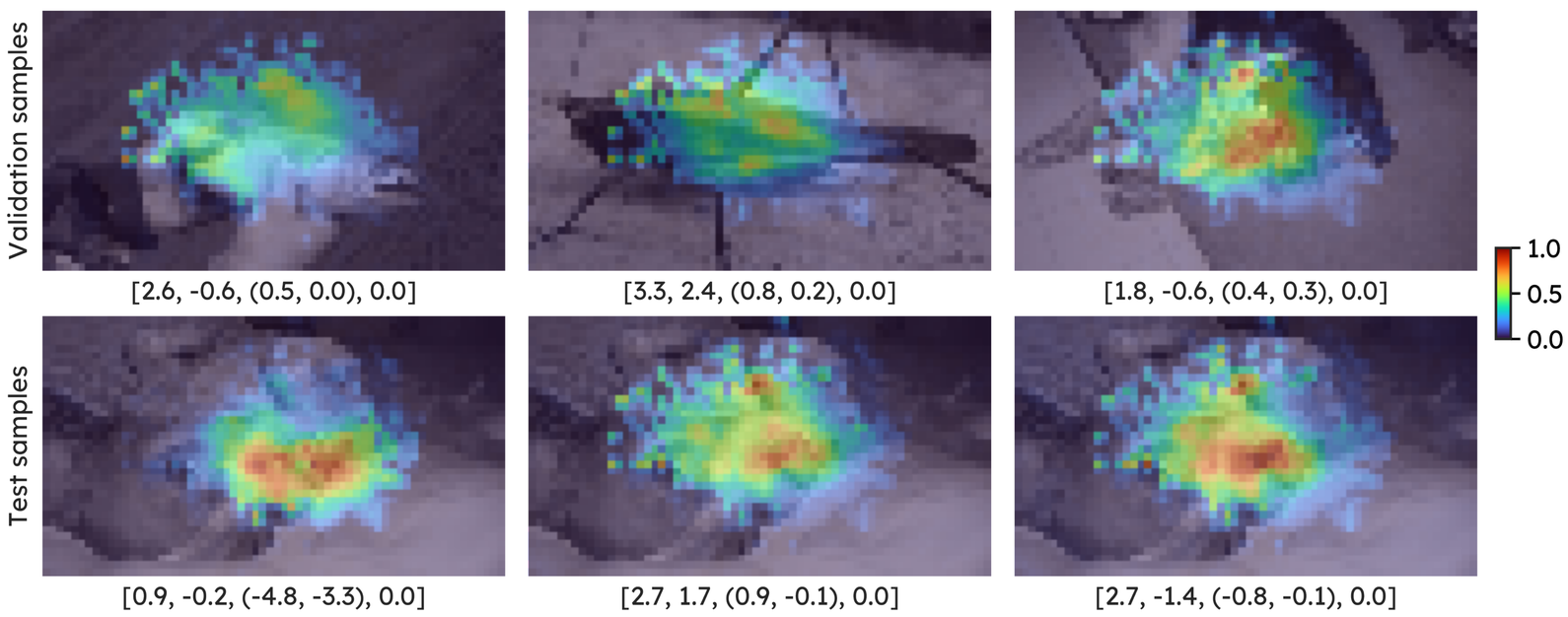
Spatial tuning difference

We evaluated the discrepancies in spatial tuning of the CNN and ViT (no behaviors) by comparing their estimated artificial receptive fields (aRFs). We found that while the two models had similar aRF locations, ViT learned notably narrower aRFs. Left: aRFs of the same unit and their Gaussian fit; Mid: Density plot of ~8k Gaussian fit centers; Right: Gaussian fit standard deviation distributions.



Self-attention visualization

We extracted the self-attention weights learned by V1T and overlay the visual stimuli from the validation (unique stimuli) and test (repeated stimuli) sets [5] with a heatmap of the learned weights. We found that the center of the self-attention maps are moderately correlated with the recorded pupil centers (horizontal: 0.525 (****), vertical: 0.409 (****)). Bracket shows the [pupil dilation, dilation derivative, pupil center (x, y), animal speed] of the trial.



Note: the term “attention” strictly refers to the self-attention layer in Transformers [4], which is distinct from the perceptual process of “attention” in the neuroscience literature.

References: [1] Bashivan et al. 2019. [2] Lurz et al. 2021. [3] Burg et al. 2021. [4] Dosovitskiy et al. 2021. [5] Willeke et al. 2022. [6] Franke et al. 2022.

github.com/bryanlimy/V1T @bryanlimy @isabelmaria_c @RochefortLab

Biomedical AI UKRI CENTRE FOR DOCTORAL TRAINING THE UNIVERSITY of EDINBURGH