From deep learning to mechanistic understanding in neuroscience: *the structure of retinal prediction*

Hidenori Tanaka,[S] Aran Nayebi,[S] N. Maheswaranathan,[G,S] Lane McIntosh,[S] Stephen A. Baccus,[S] Surya Ganguli[S,G]
Stanford University,[S] Google Brain,[G] NTT Physics & Informatics Lab

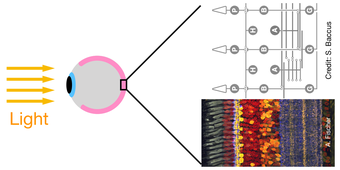
What can deep learning models tell us about the inner-workings of the brain?

We combine "*in-silico* neurophysiological experiments" and "interpretation tools" to perform model reduction, yielding new experimentally testable mechanisms.

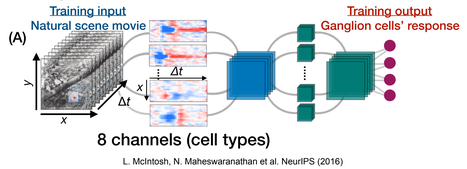
Introduction

Deep learning models are successful at reproducing the input-output map of sensory areas (retina, V1, V4, IT). However, can these models generate experimentally testable mechanistic hypotheses of internal biological computations?
We show that model reduction on artificial networks can generate new experimentally testable hypotheses.

- Deep learning mostly concerns the accuracy of the input-output mapping.
- Neuroscience aims to understand the internal computational mechanisms of generating the output given sensory stimuli.



Training procedure: natural scenes

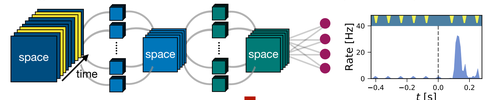


L. McIntosh, N. Maheswaranathan et al. NeurIPS (2016)

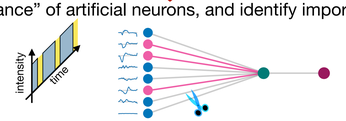
Model reduction: Identifying *important* sub-circuits

Average over uniform directions and keep only important units

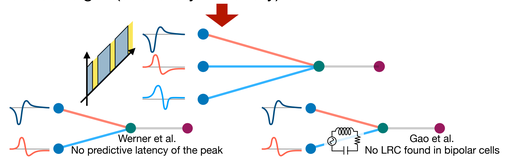
Average over uniform directions of artificial spatiotemporal stimuli



Quantify "importance" of artificial neurons, and identify important sub-circuits



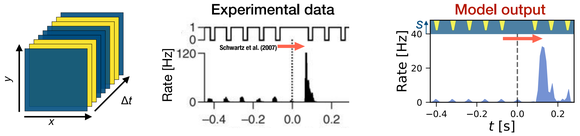
chose "sign" (excitatory/inhibitory) of artificial neurons



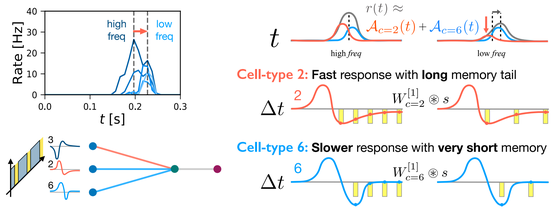
Compare with existing circuit models in systems neuroscience

A "new and only" model that captures "predictive latency"

One ON-bipolar pathway is always active, while the other is only active in high-frequency



Q1. What computational mechanism causes the large amplitude burst?
Q2. How is the latency of the peak proportional to the period of the flashes?

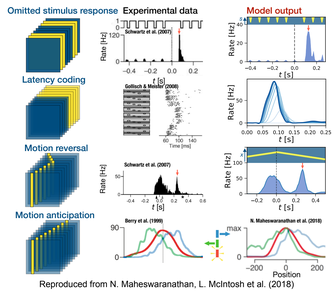


Higher frequency → **Cell-type 2** responds more strongly → shifts response earlier

Are the artificial network's computational mechanisms for generating neural responses the same as those in the brain?

Can we bridge the Gap?

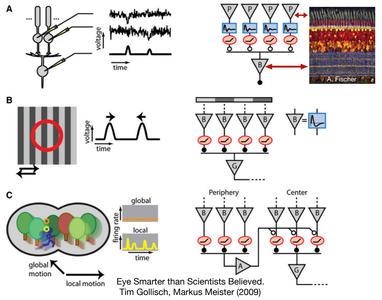
Prediction by deep learning models



Reproduced from N. Maheswaranathan, L. McIntosh et al. (2018)

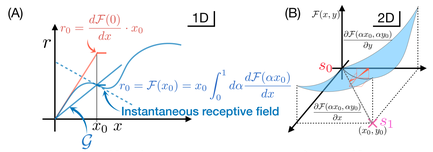
Convolutional neural network models
- originally inspired by biological neural circuits
- potentially over-parameterized
- generalize to wide arrays of visual stimuli

Models in systems neuroscience literatures



Circuit models in systems neuroscience
- Consists of biological building blocks
- concise mathematical description
- testable by neurophysiology experiments

Integrated Gradients: How important is a neuron?



$$\mathcal{F}(\mathbf{s}) = \sum_{i=1}^{N}\left(\int_0^1 d\alpha \left.\frac{\partial \mathcal{F}}{\partial s_i}\right|_{\mathbf{s}(\alpha)} s_i(\alpha)\right) = \sum_{i=1}^{N} A_i$$

output, stimulus

M. Sundararajan et al. ICML (2017)

(i) Apply chain rule to evaluate importance of hidden layer neurons

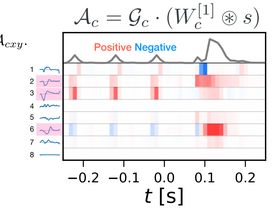
$$r(t) = \sum_{x,y,c}\left[\int_0^1 d\alpha \left.\frac{\partial \mathcal{F}}{\partial z^{[1]}_{cxy}}\right|_{s(t,\alpha)}\right](W^{[1]}_{cxy} \circledast s) = \sum_{x,y,c}[\mathcal{G}_{cxy}(s)]\,(W^{[1]}_{cxy} \circledast s) = \sum_{x,y,c}\mathcal{A}_{cxy}.$$

channel, filter, stimulus

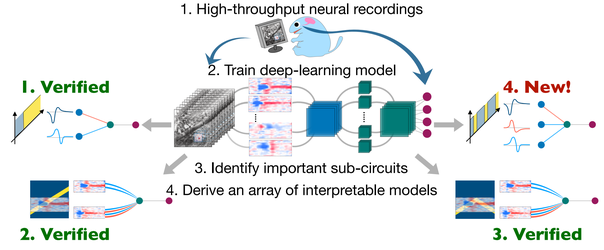
(ii) Exploit stimulus invariances to reduce dimensionality.

$$r(t) = \sum_{c=1}^{8}\left(\sum_{x=1}^{36}\sum_{y=1}^{36}\mathcal{G}_{cxy}(s)\right)\cdot(W^{[1]}_{c} \circledast s) \equiv \sum_{c=1}^{8}\mathcal{G}_c(s)\cdot(W^{[1]}_{c} \circledast s) \equiv \sum_{c=1}^{8}\mathcal{A}_c.$$

$$\mathcal{A}_c = \mathcal{G}_c \cdot (W^{[1]}_c \circledast s)$$



Model reduction on artificial network yielded three experimentally verified mechanisms and a new testable hypothesis.



1. Latency coding: Dual ON-OFF bipolar pathways model consistent with (i) pharmacological experiments, (ii) existing theory

2. Motion anticipation: Dual pathways are necessary and inhibitory cell type is necessary.
Qualitatively consistent with a recently proposed theory.

3. Motion reversal: ON/OFF reversal along space explains existence of the burst. Inhibitory pathway explains the fixed latency. Qualitatively consistent with a recently proposed theory

4. Omitted Stimulus Response: Derived a new & only theory that agrees with all experimental facts. The model predicts existence of two types of ON bipolar cells (one only active in high frequency regime with earlier peak, the other always active with later peak) both sending inputs to the same ganglion cell.