ASHOKA UNIVERSITY

Investigating Brain-like CNNs and Consequences

IICCSSS Blitz Talk by **Niranjan Rajesh**

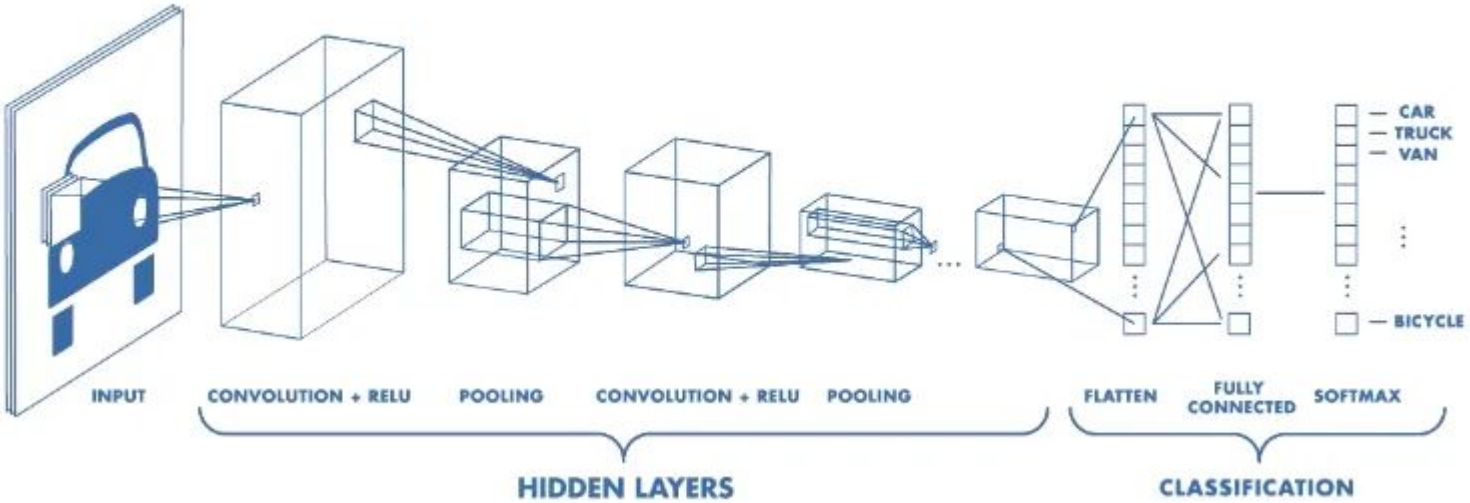
Contents

- Background
- Brain-like CNNs
- Adversarial Robustness
- Neural Manifolds

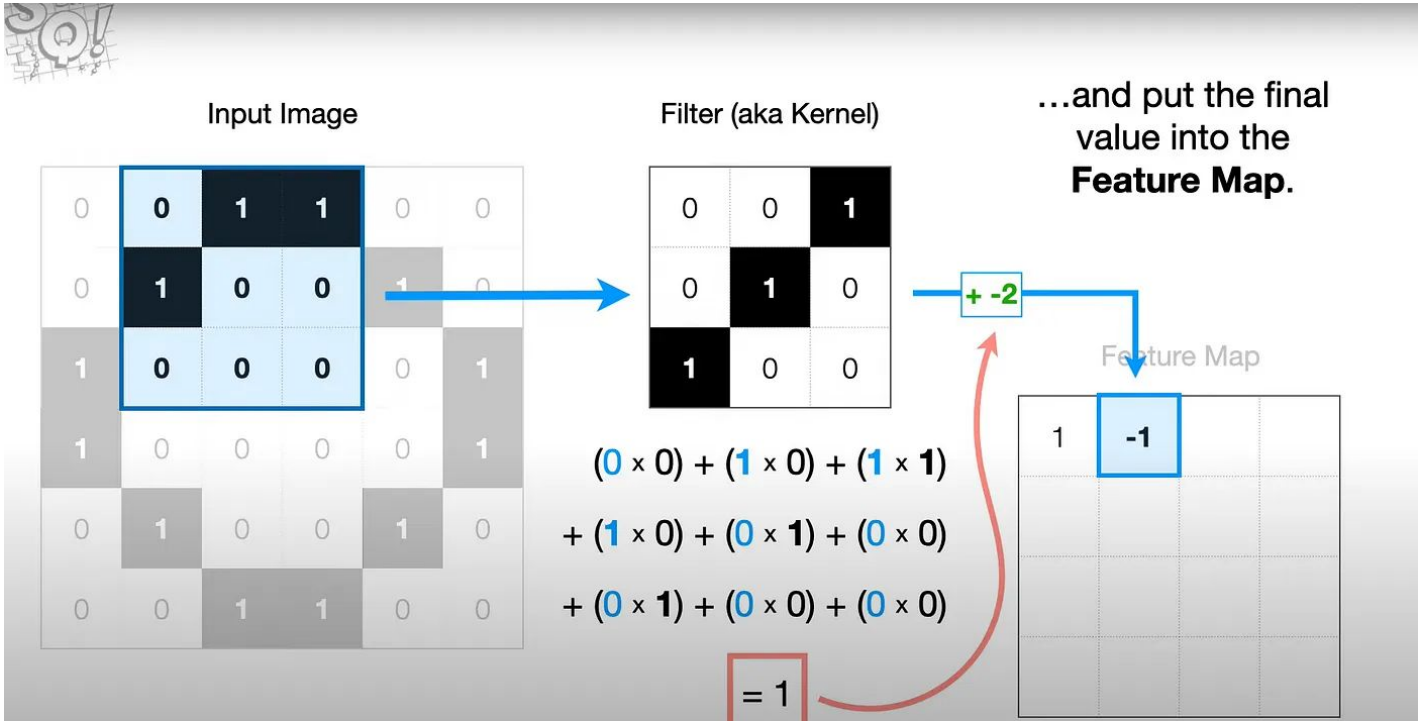
Some Background

CNNs and their problems

What is a CNN?

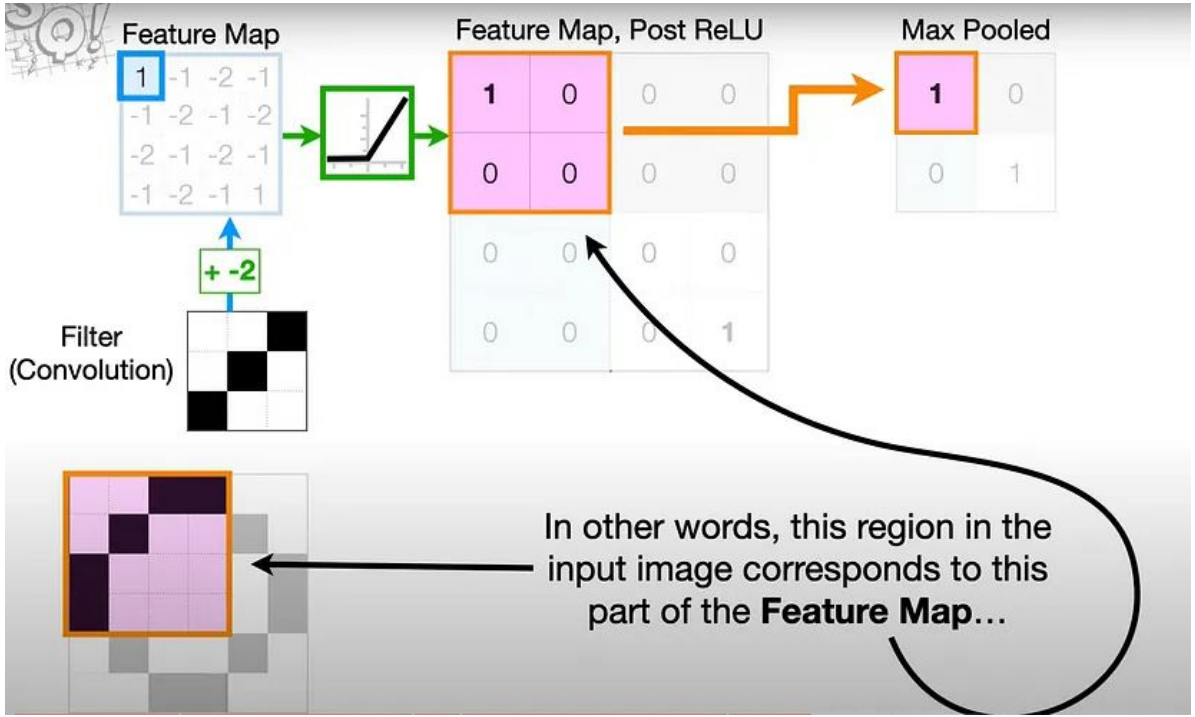

CAR
TRUCK
VAN
BICYCLE
INPUT
CONVOLUTION + RELU
POOLING
CONVOLUTION + RELU
POOLING
FLATTEN
FULLY CONNECTED
SOFTMAX
HIDDEN LAYERS
CLASSIFICATION

Filters and Feature Maps



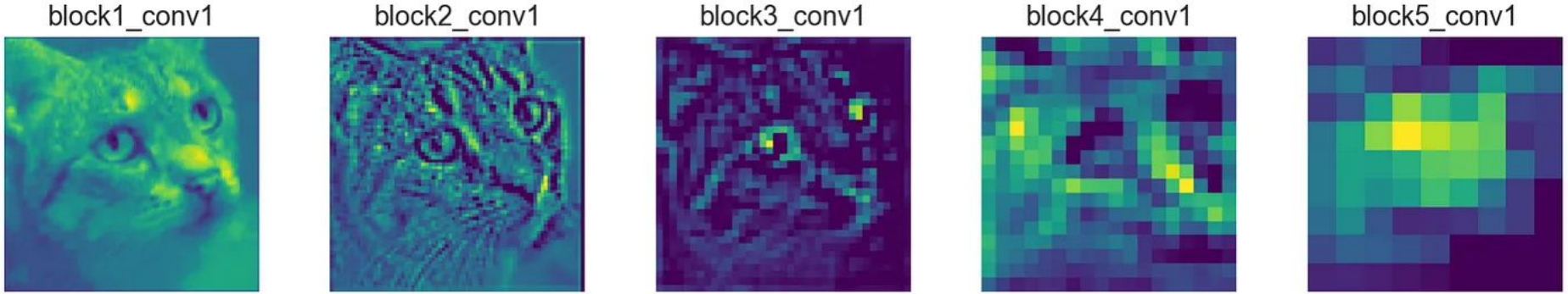
(StatQuest)

Non-linearity and Pooling



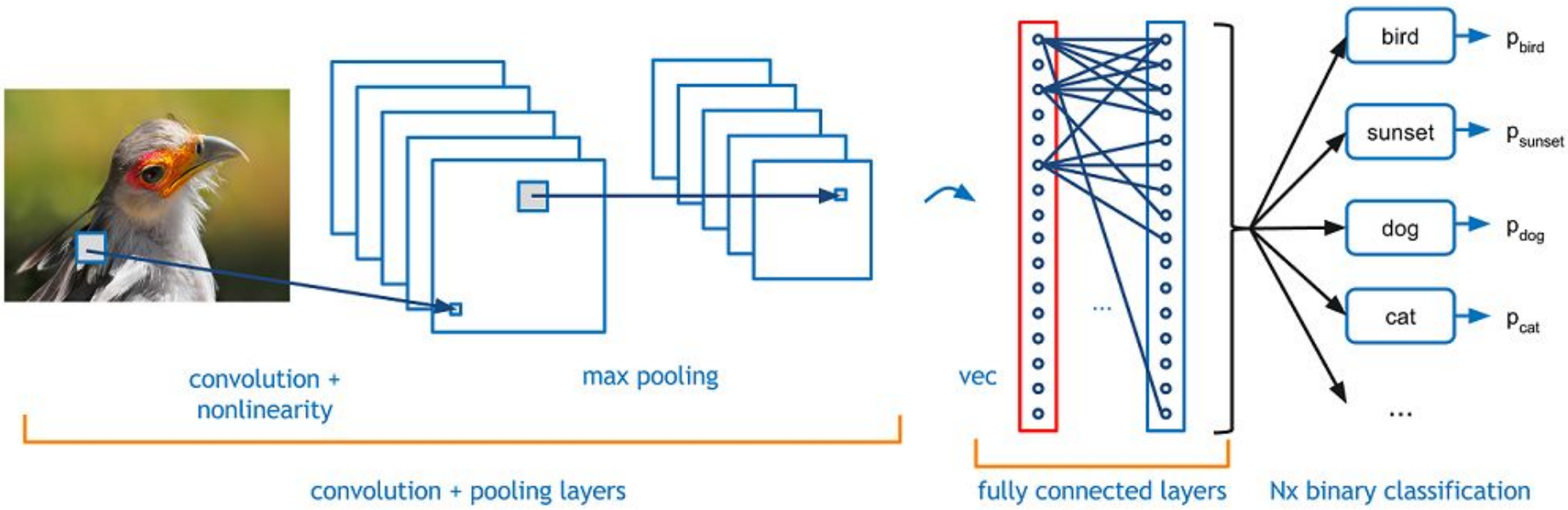
(StatQuest)

Example Feature maps



(Towards Data Science)

Finally


convolution +
nonlinearity
max pooling
vec
bird
sunset
dog
cat
...
p_{bird}
p_{sunset}
p_{dog}
p_{cat}
convolution + pooling layers
fully connected layers
Nx binary classification

Question

How is CNN vision different from humans?

Semantic Understanding

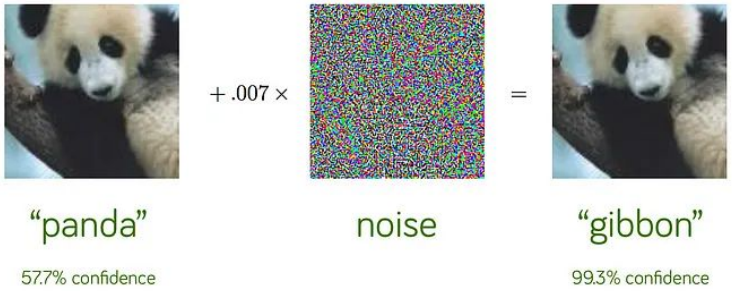
Question

How is CNN vision different from humans?

Generalisation

Question

How is CNN vision different from humans?



Adversarial Robustness

Question

How is CNN vision different from humans?

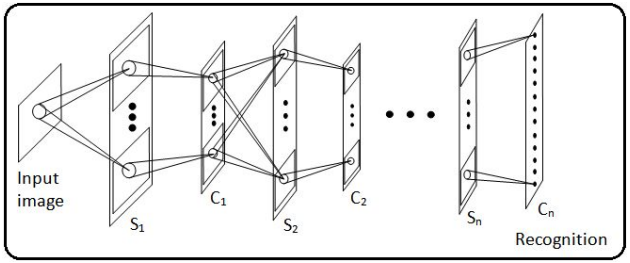
VERY different

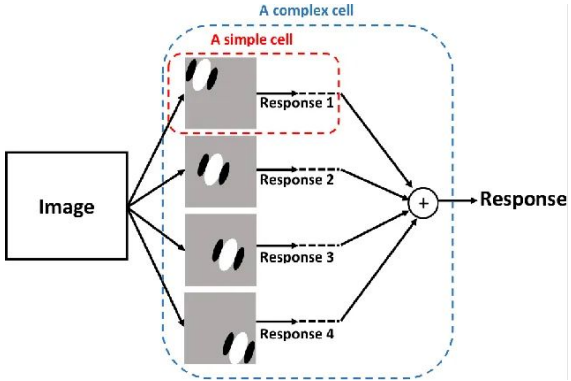
Brain-like CNNs

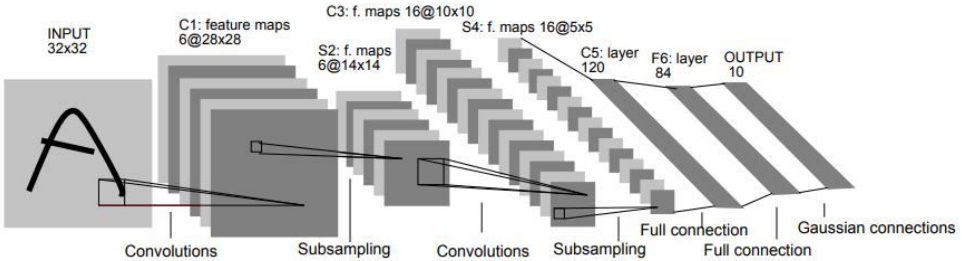
Bridging the gap?

Historical inspiration from Neuroscience

- Hubel and Wiesel (1959) - Simple and Complex cells in the Cat Striate Cortex
- Fukushima (1980) - the Neocognitron
- LeCun (1989) - LeNet, the first ‘convolutional’ neural network
- AlexNet, VGGNet, ResNets, …

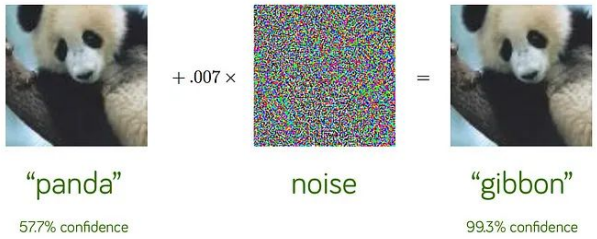






Motivation for further alignment

- Huge gap between the performance of primate visual processing and state-of-the-art CNNs
- Models seem to be learning differently to primates - space for improvement
- Can we turn back to neuroscience?

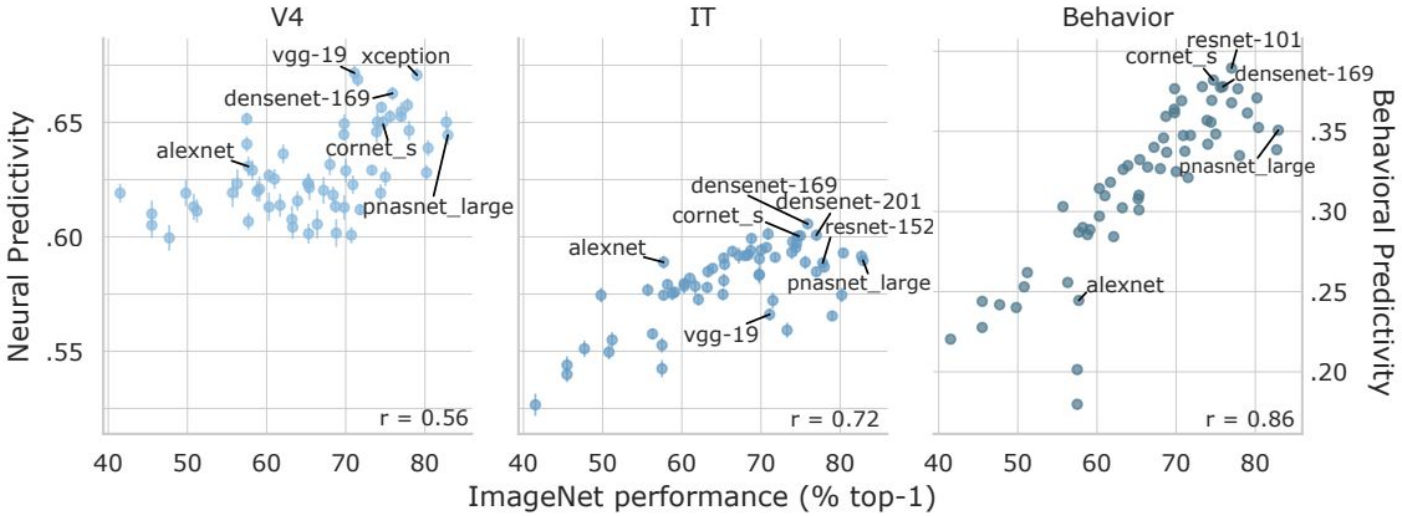


Benchmarking Neural Similarity

- Brain-Score (Schrimpf et al, 2020) - a platform for neural and behavioural similarity benchmarks for Neural Networks
- How well do the internal representation of the network predict the internal representations of the primate visual system? **Neural Predictivity**
- How different is the *output* of the network to that of the primate visual system (erroring, etc.)? **Behavioural Similarity**

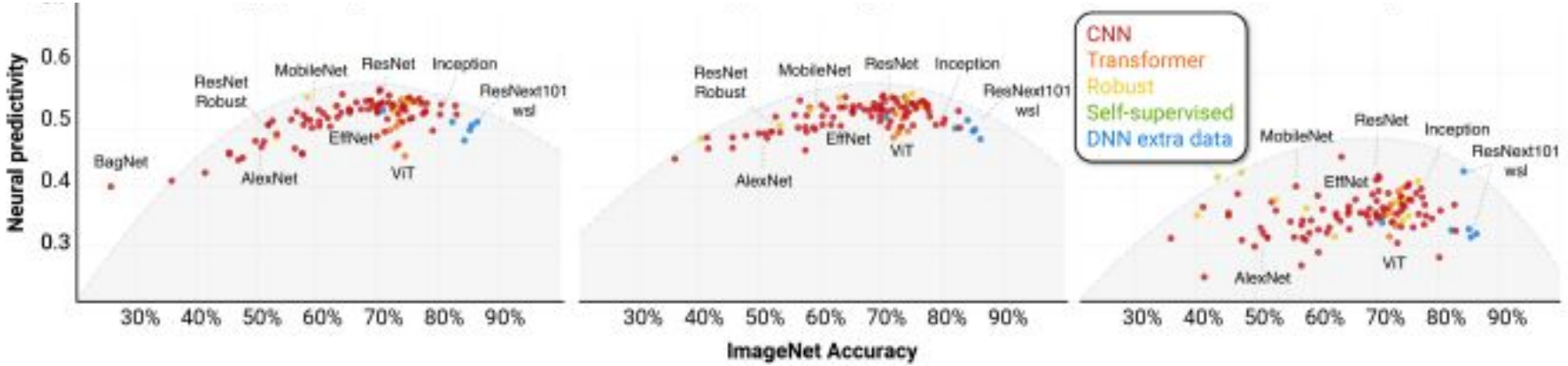
(Schrimpf et al, 2020)

Benchmarking Neural Similarity



(Schrimpf et al, 2020)

Benchmarking Neural Similarity



(Linsley et al, 2023)

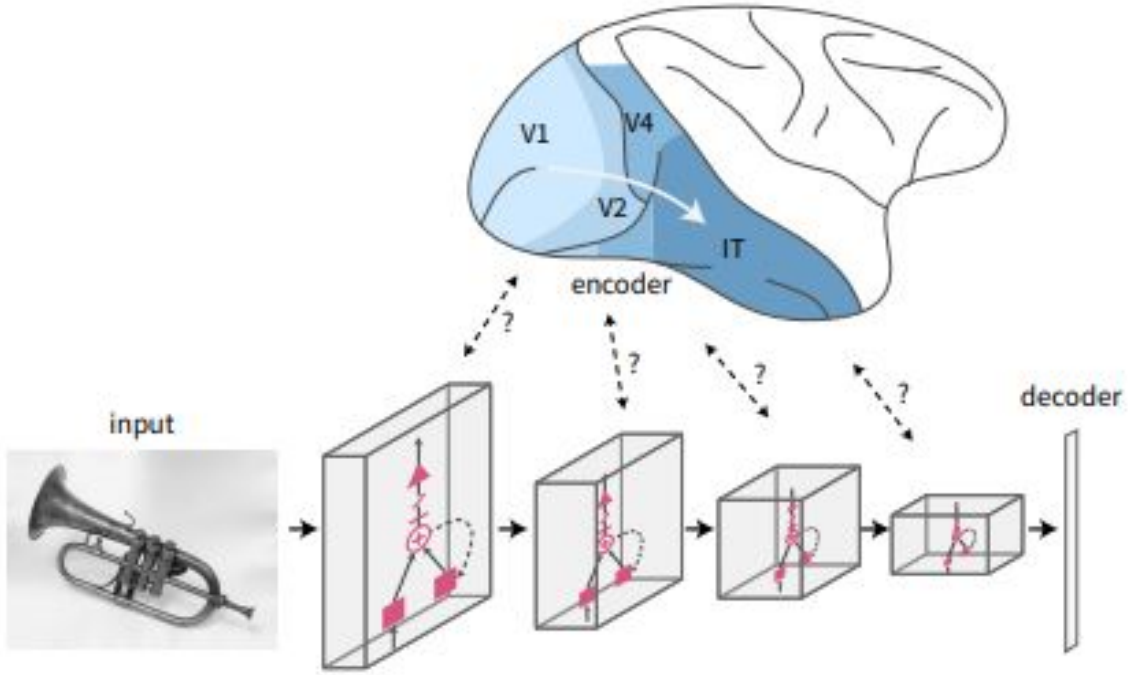
Can models capture brain process similarity more stringently and not as a coincidence?

Mapping layers to brain regions

- What if we align a network to the neuroanatomy of a brain? **CORnet-S**
- CORnet-S criteria:
  - Predictivity
  - Compactness
  - Recurrence
- An honest and mechanistic model of the brain

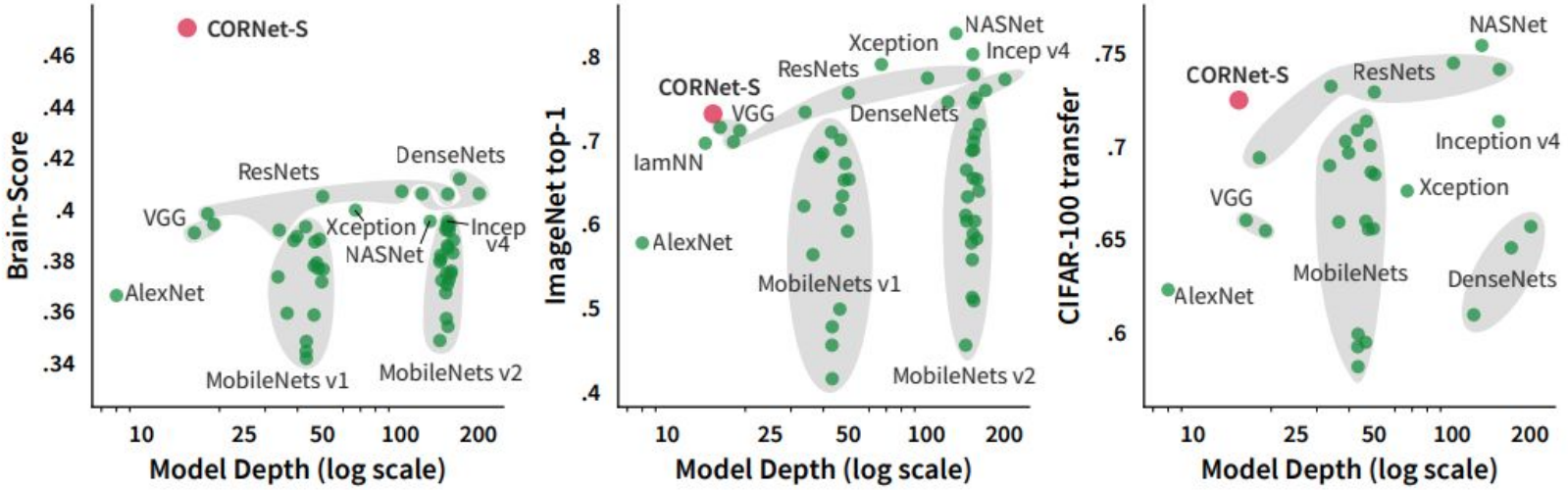
(Kubilius et al, 2019)

Mapping layers to brain regions



(Kubilius et al, 2019)

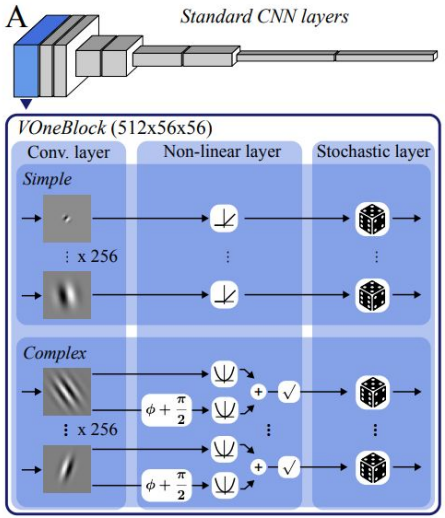
Mapping layers to brain regions



(Kubilius et al, 2019)

Architectural constraints

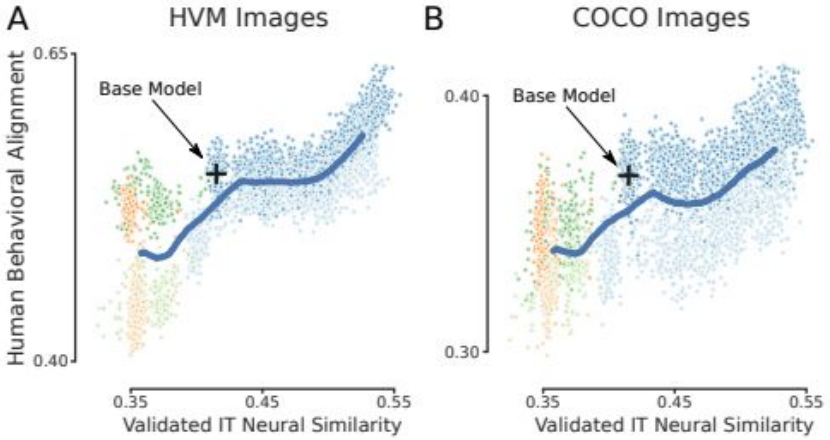
- What if we ‘fit’ classical neuroscientific models of primate visual regions into CNNs - **VOneNet**
- Modelled after the Linear-Nonlinear Poisson (LNP) Model of V1
- Approximates primate neural processing of images better than SOTA counterparts



(Dapello et al, 2020)

Representational Constraints

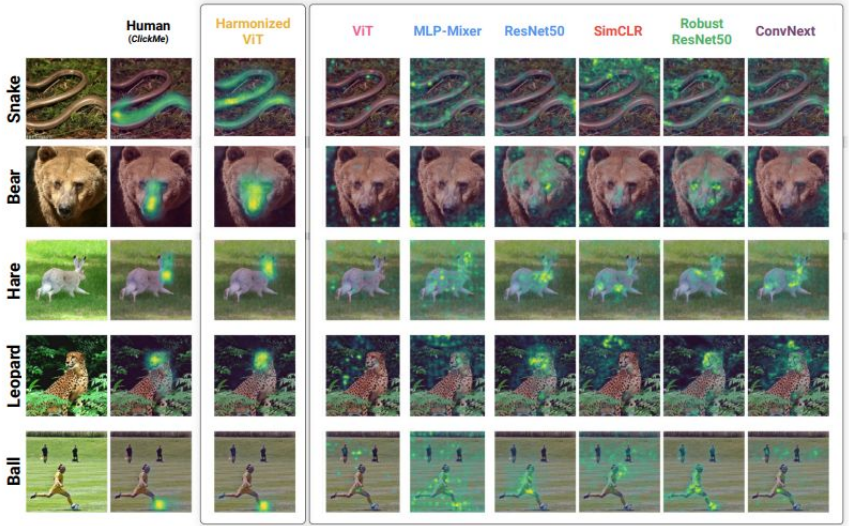
- What about simulating the architecture, the representational output of the systems are matched?
- Multi-loss setup:
  - Standard Categorical Cross-Entropy for ImageNet Classification
  - Centered Kernel Alignment (CKA) loss for ‘IT’ layer representation alignment



(Dapello et al, 2023)

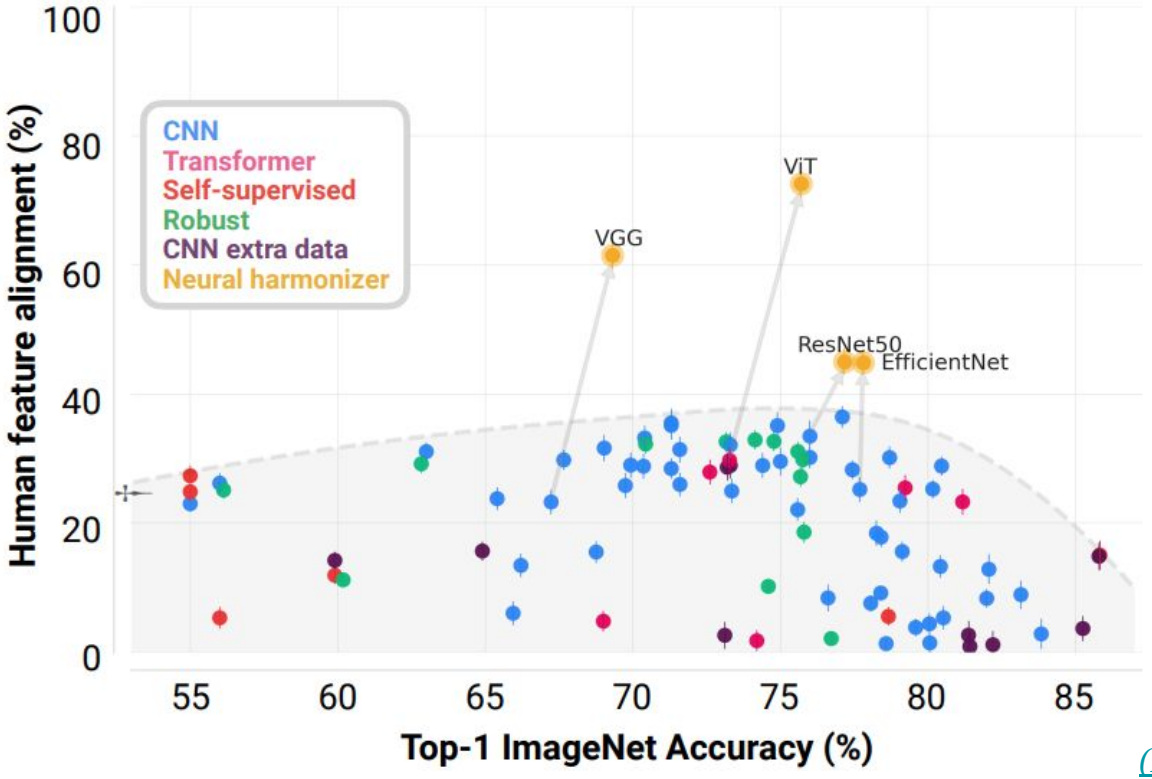
Behavioural Constraints

- What about alignment at the ‘final’ level - behaviour
- Aligning feature importance maps - **Neural Harmonisation**



(Fel et al, 2022)

Behavioural Constraints

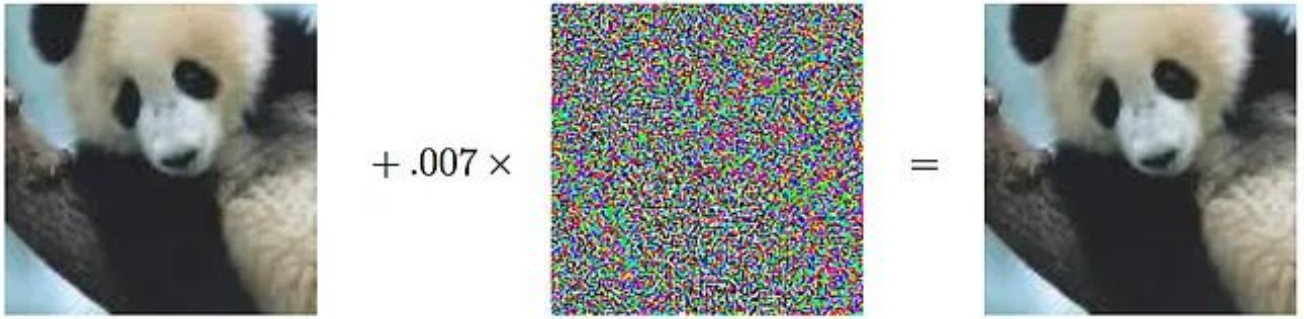


(Fel et al, 2022)

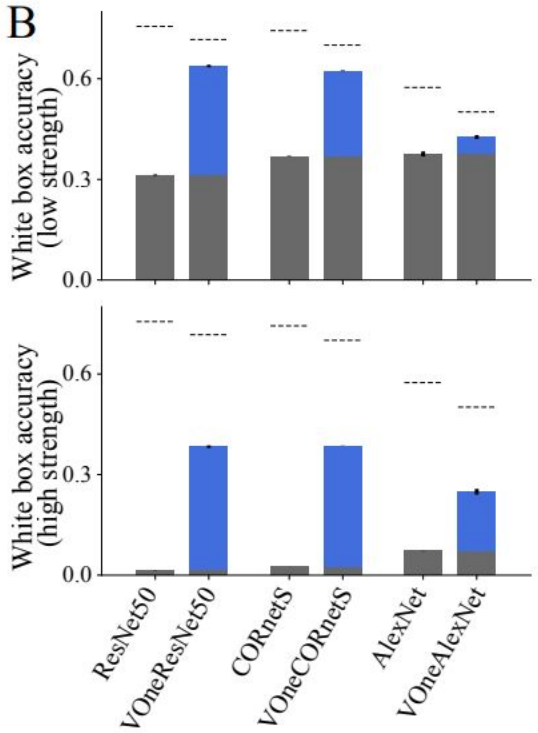
Adversarial Robustness

All these different constraints but one common consequence

Adversarial Attacks

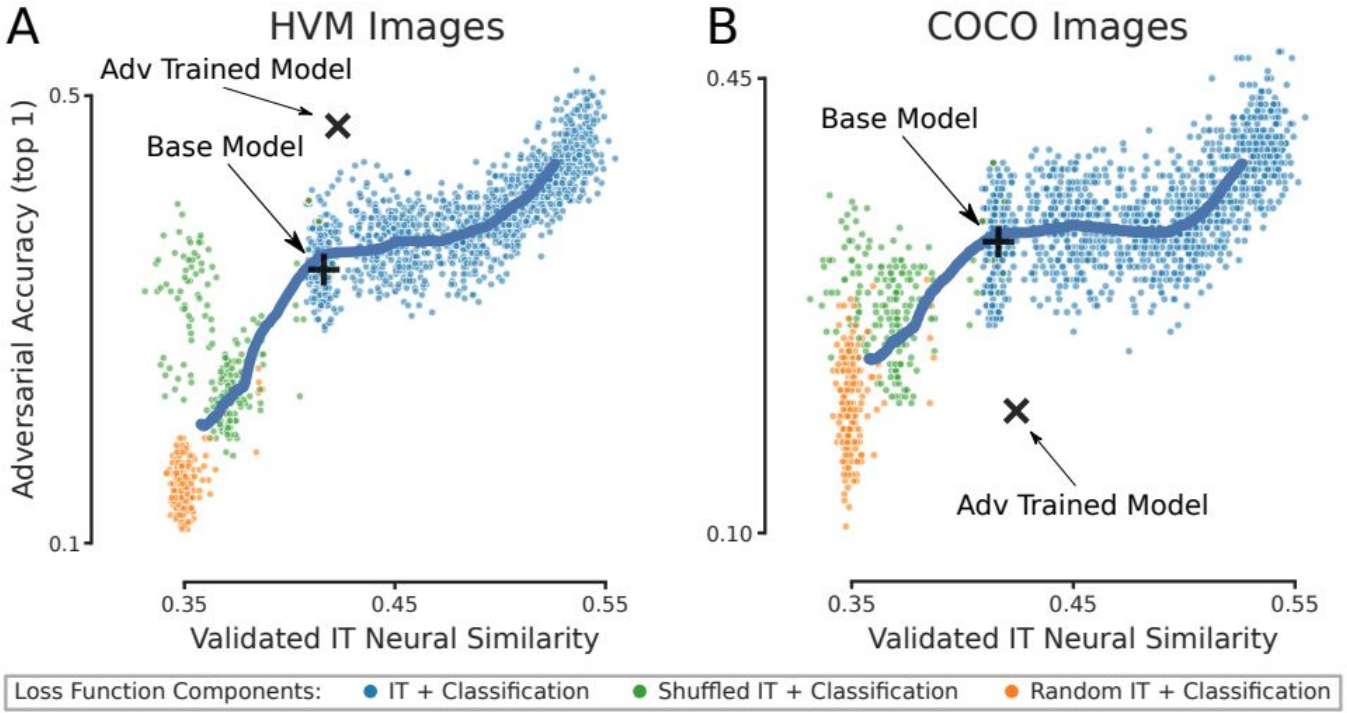


Effects on Robustness



(Dapello et al, 2020)

Effects on Robustness



(Dapello et al, 2023)

Effects on Robustness

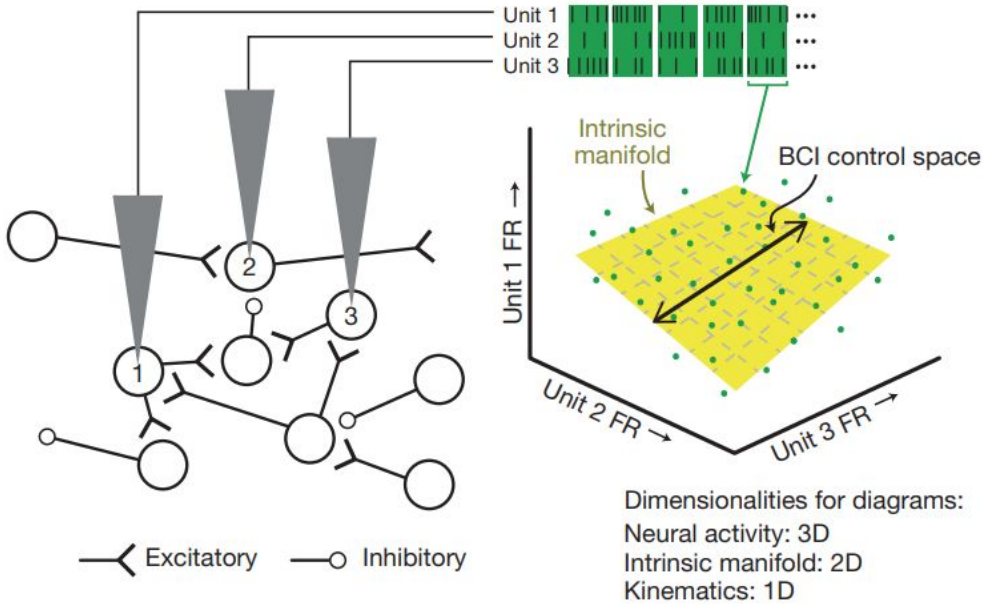
- Robustness to such attacks was not explicitly coded for. How did they arise?
- “Easy - they are made to be more like the animal brain so of course they behave more like the animals in visual tasks”
- But **how?** What exactly is in the ‘neural’ code that achieves adversarial robustness?

Neural Manifolds

The first piece of the puzzle?

What are Neural Manifolds

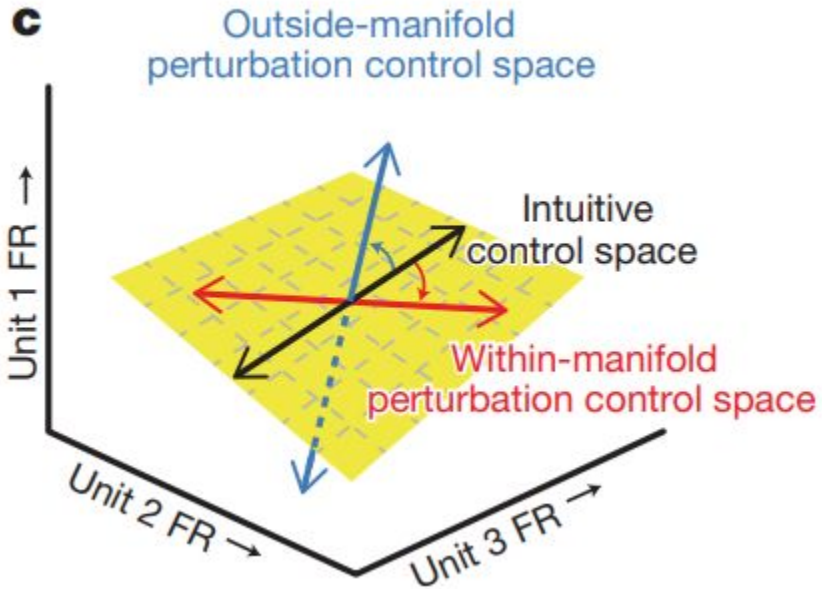
- Task-specific subspaces within the larger neural space



(Sadtler et al, 2014)

Within and outside manifold tasks

We can manipulate the neural space and investigate the effects on task performance



(Sadtler et al, 2014)

Investigation:

- Identification of CNN manifold through neural unit representations
- Compute higher dimensional ‘positions’ of representations of varied input within manifolds
- Manipulation of the neural space and manifold

Can we verify if the adversarial attacks are within the brain-nets’ manifolds?

Implications:

- Manifold hypothesis holds some water
- Adversarial Robustness through manifold engineering
- Other ‘traits’ or skills?

Thank you

niranjanrajesh02@gmail.com
niranjan.rajesh_asp24@ashoka.edu.in