Does Leveraging the Human Ventral Visual Stream Improve Neural Network Robustness?

Zhenan Shao[1], Linjian Ma[2], Bo Li[2,3], Diane M. Beck[1]

BIAI 03/21/2024

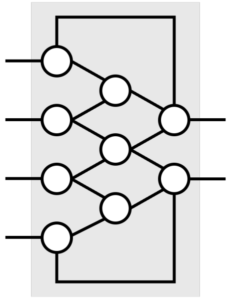
[1]Department of Psychology, UIUC
[2]Department of Computer Science, UIUC
[3]Department of Computer Science, UChicago

Introduction

Robust human vision and vulnerable machine vision

Human visual perception achieves various invariance

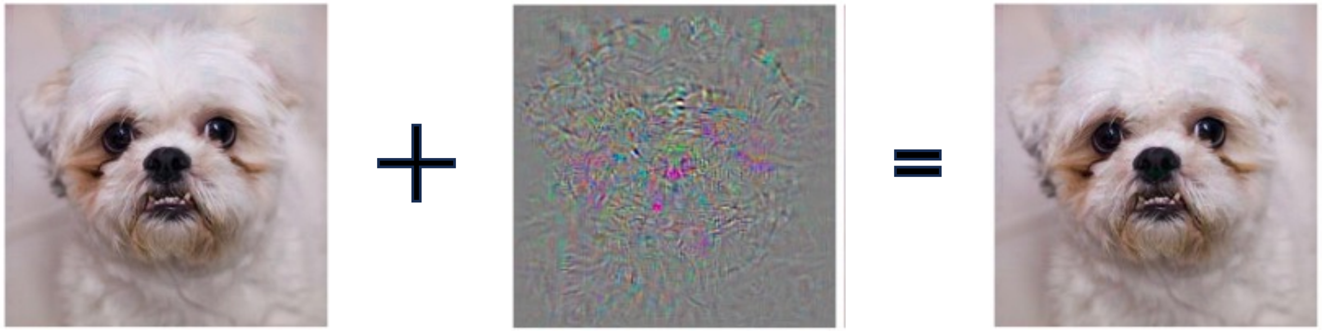
Biederman & Cooper, 1991;
Cave, Bost & Cobb, 1996;
Biederman & Gerhardstein, 1993;
…



Even imperceptible perturbations can lead to wrong prediction

Szegedy et al., 2014;
Carlini & Wagner, 2017;
Kurakin et al., 2017;
…

dog　　　　　　　　　　Ostrich

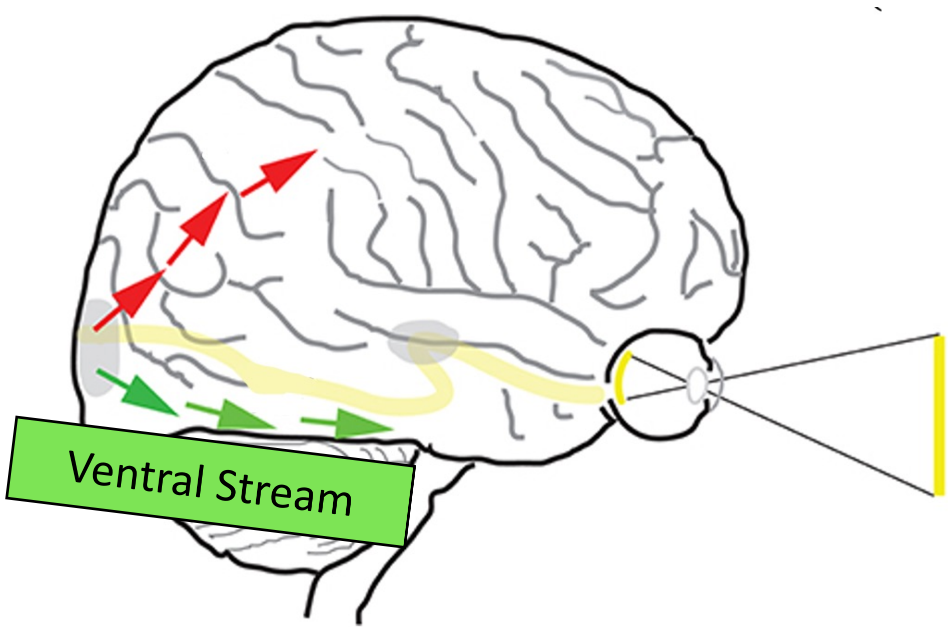


(Szegedy et al., 2014)

Introduction

Achieving invariances along visual ventral stream

- The ventral visual stream forms a hierarchy, transitioning from basic visuals to more abstract and stable representations (Logothetis and Sheinberg, 1996, Zoccolan et al., 2007, Isik et al., 2014, Iordan et al., 2015…).
- Evolving representations space achieved by separating object manifolds (Dicarlo & Cox, 2007).

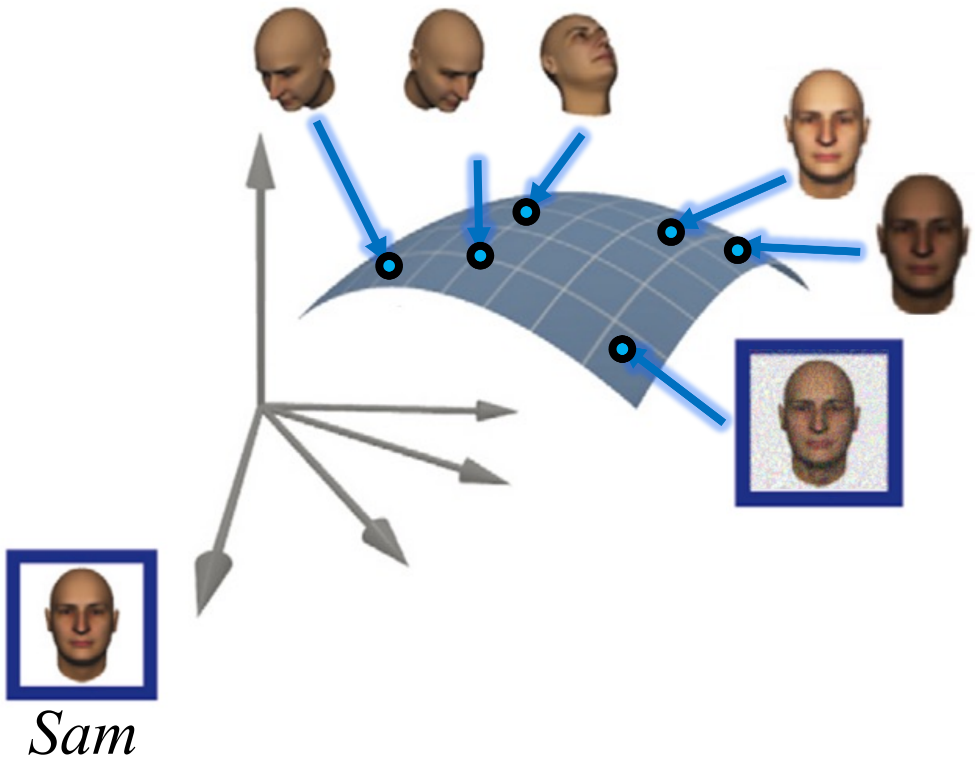


(Sheth & Young, 2016)

Introduction

Achieving invariances along visual ventral stream

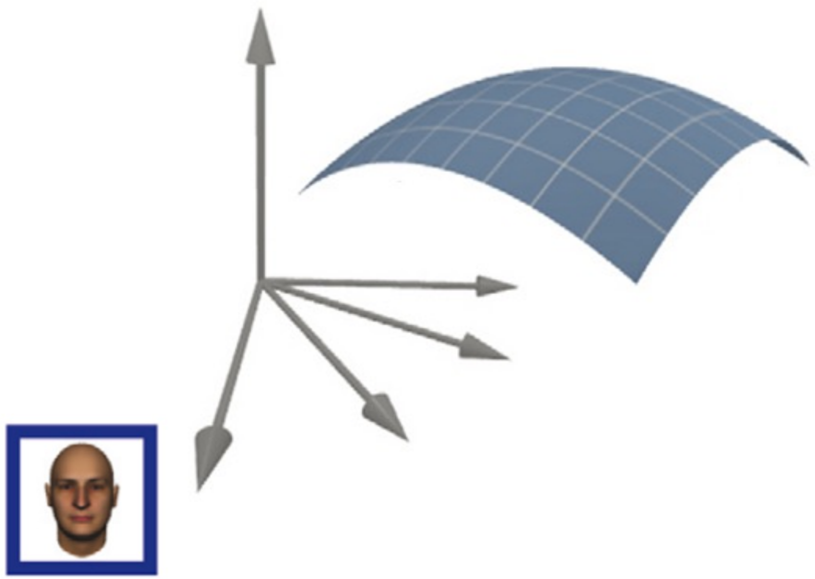
- The ventral visual stream forms a hierarchy, transitioning from basic visuals to more abstract and stable representations (Logothetis and Sheinberg, 1996, Zoccolan et al., 2007, Isik et al., 2014, Iordan et al., 2015…).
- Evolving representations space achieved by separating object manifolds (Dicarlo & Cox, 2007).



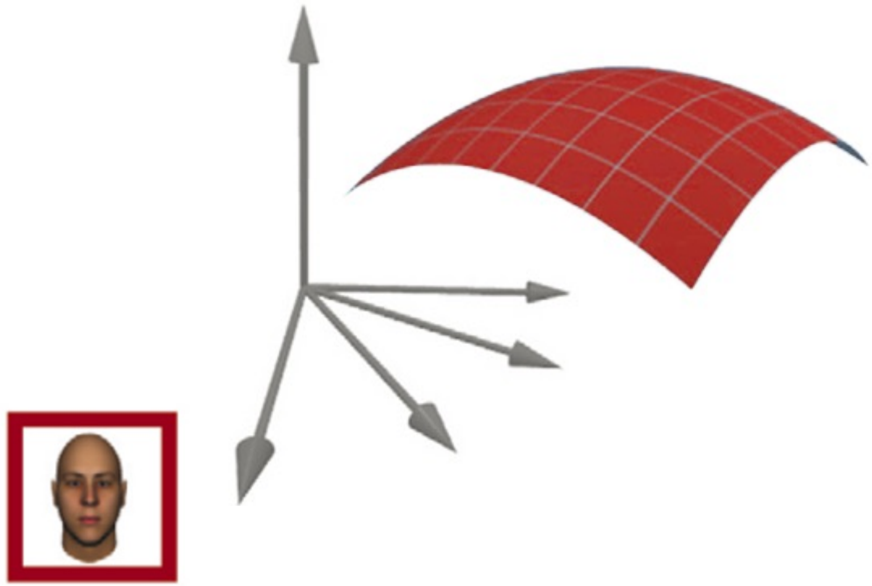
Introduction

Achieving invariances along visual ventral stream

- The ventral visual stream forms a hierarchy, transitioning from basic visuals to more abstract and stable representations (Logothetis and Sheinberg, 1996, Zoccolan et al., 2007, Isik et al., 2014, Iordan et al., 2015…).
- Evolving representations space achieved by separating object manifolds (Dicarlo & Cox, 2007).



Sam

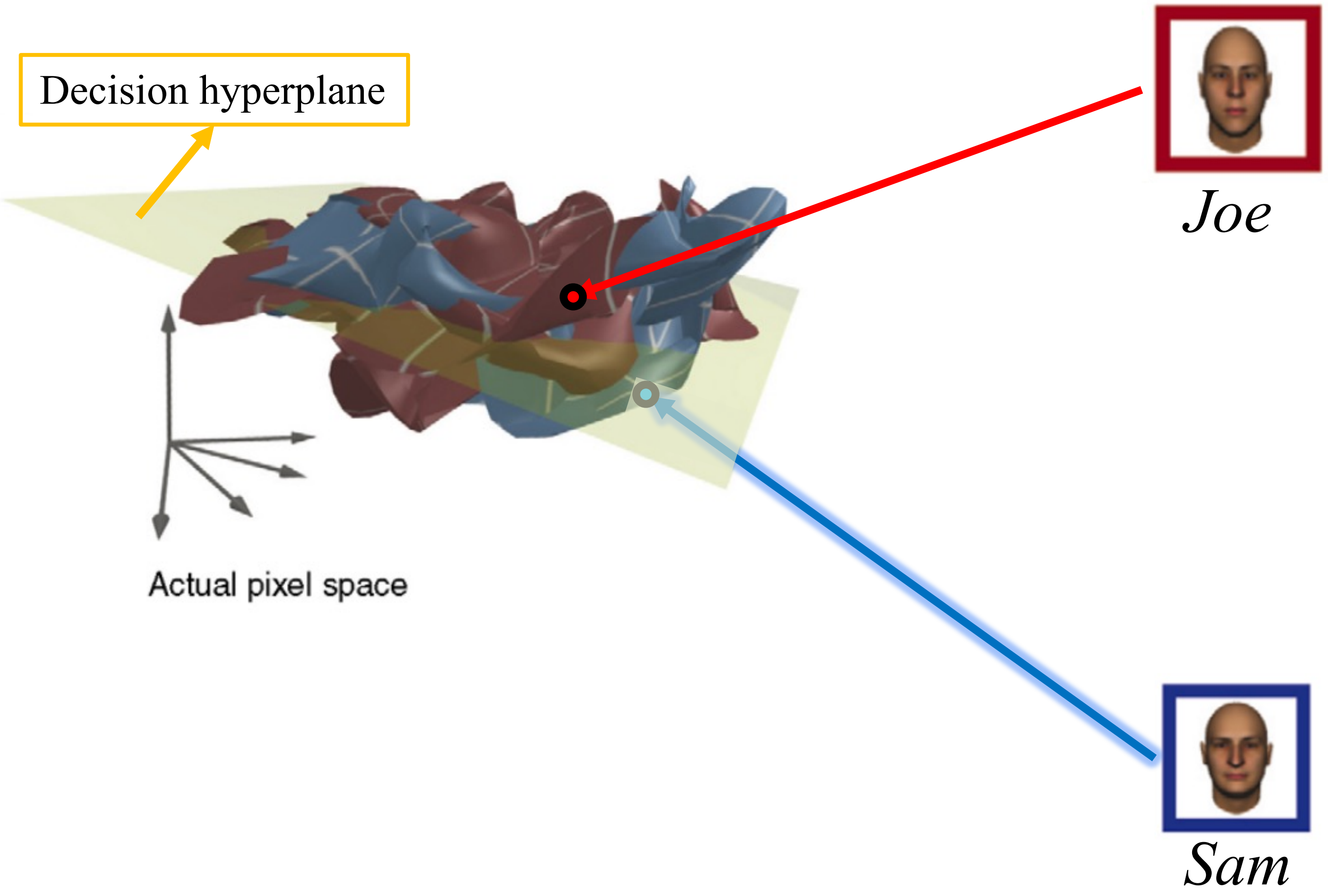


Joe

Introduction

Achieving invariances along visual ventral stream

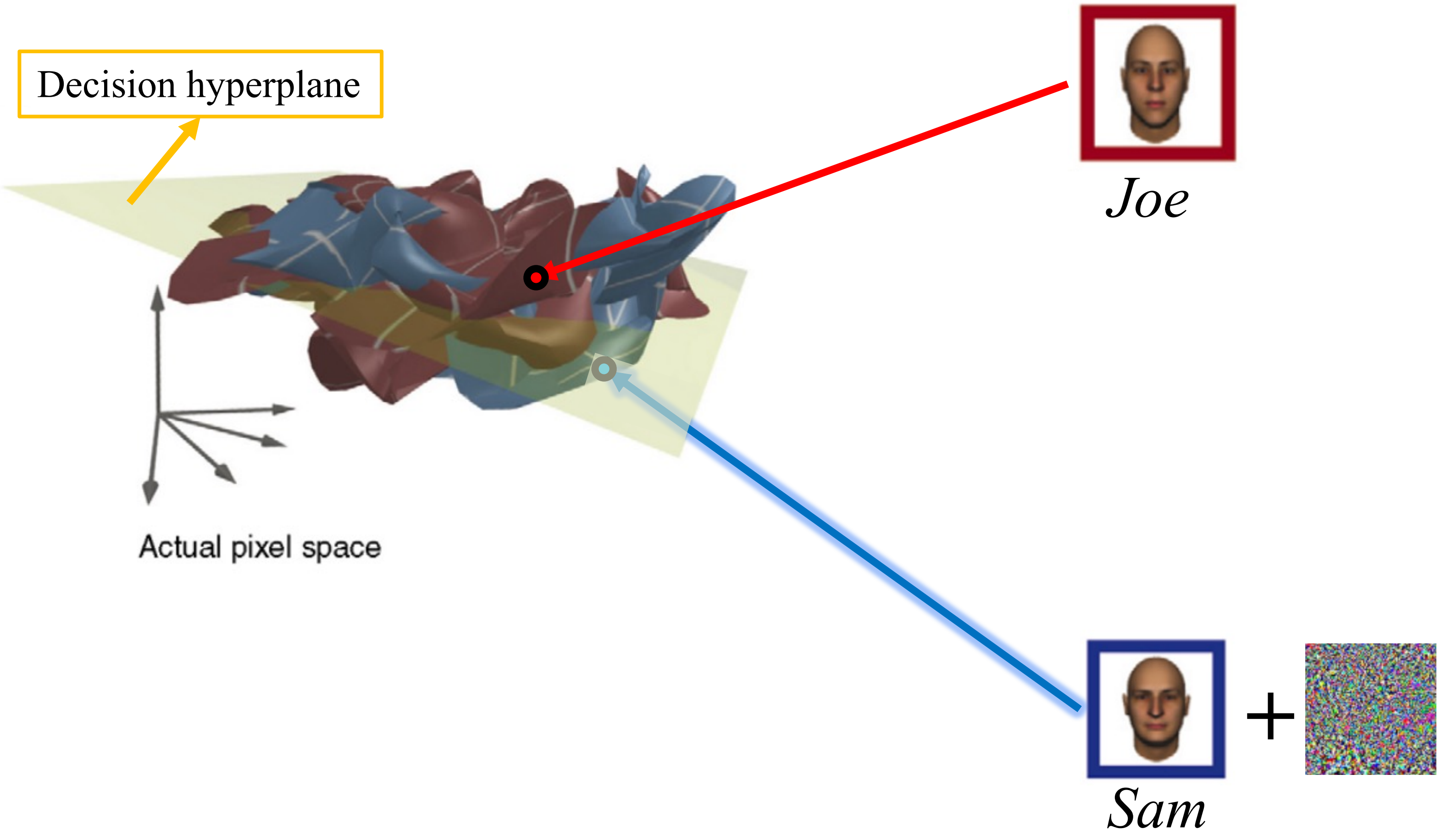
- Evolving representations achieved more separable object manifolds (Dicarlo & Cox, 2007).



Introduction

Achieving invariances along visual ventral stream

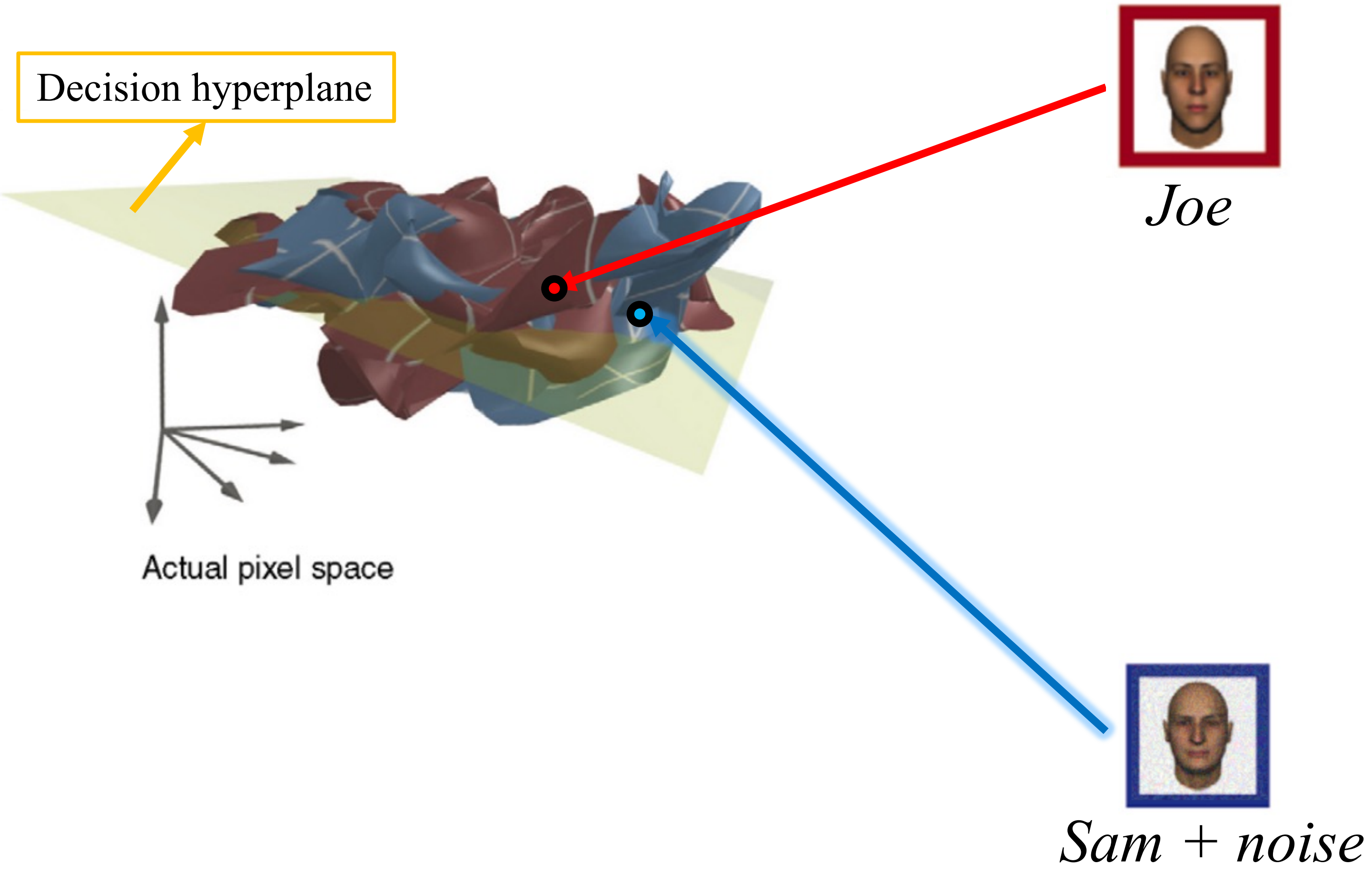
- Evolving representations achieved more separable object manifolds (Dicarlo & Cox, 2007).



Introduction

Achieving invariances along visual ventral stream

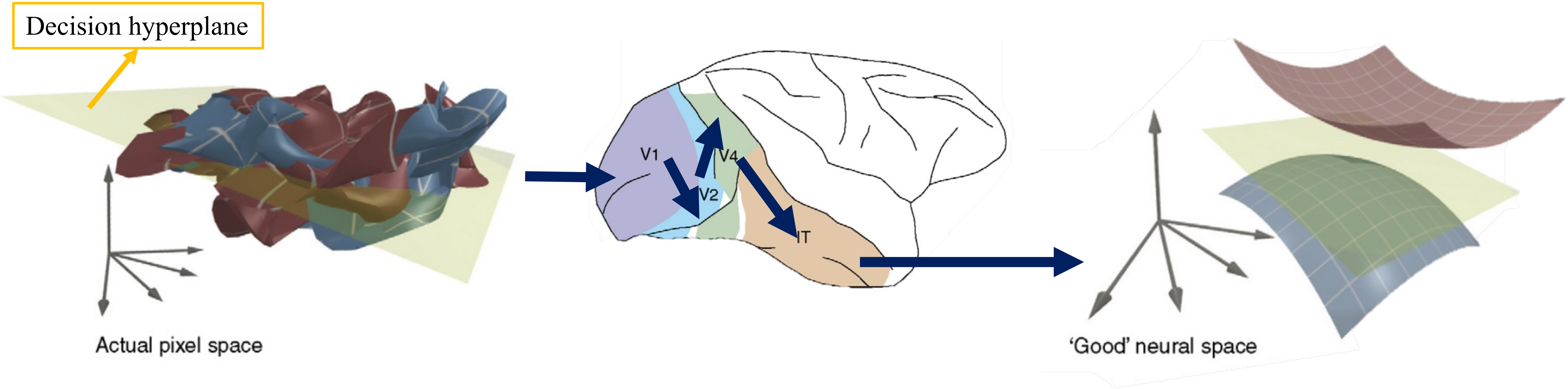
- Evolving representations achieved more separable object manifolds (Dicarlo & Cox, 2007).



Introduction

Achieving invariances along visual ventral stream

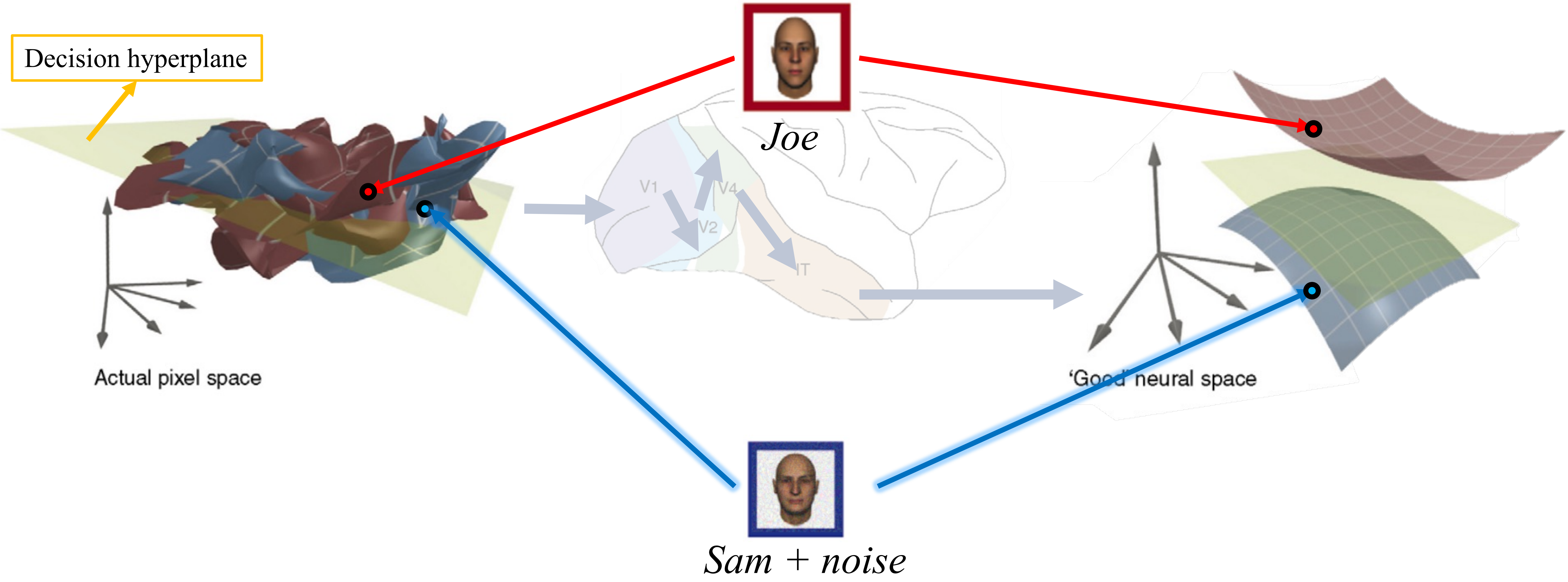
- Evolving representations achieved more separable object manifolds (Dicarlo & Cox, 2007).



Introduction

Achieving invariances along visual ventral stream

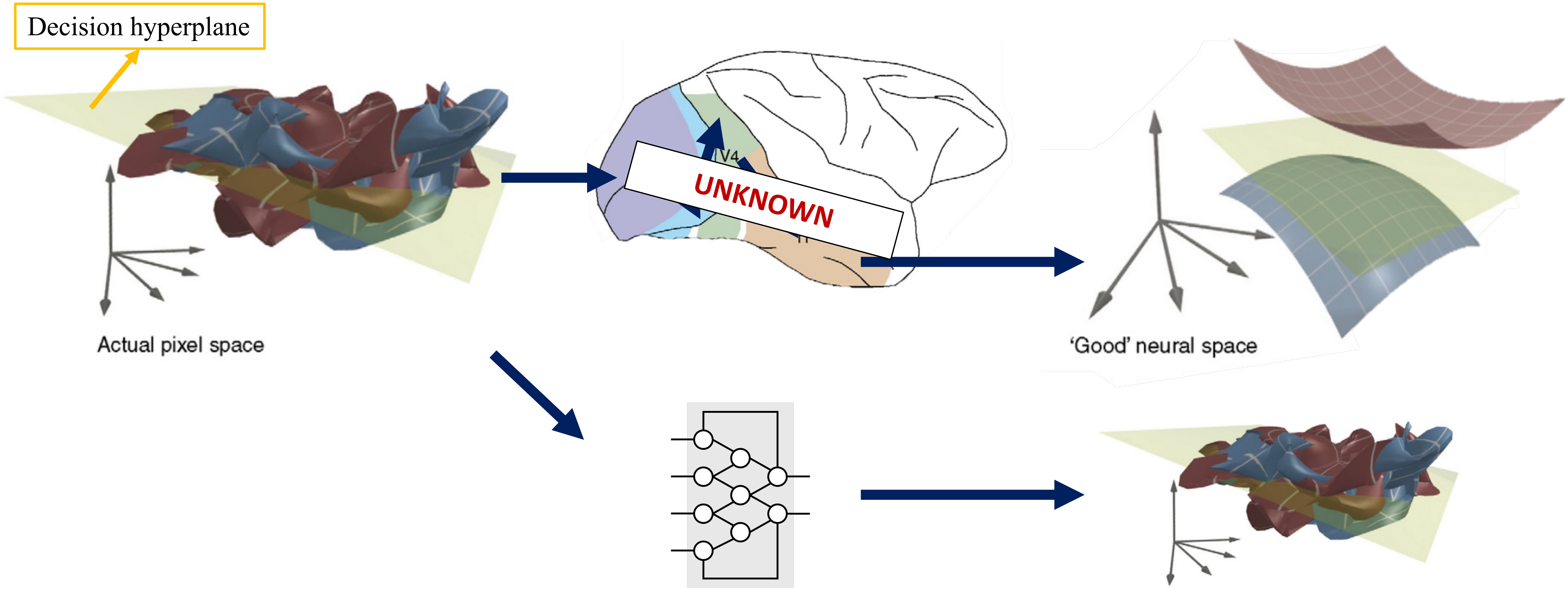
- Evolving representations achieved more separable object manifolds (Dicarlo & Cox, 2007).



Introduction

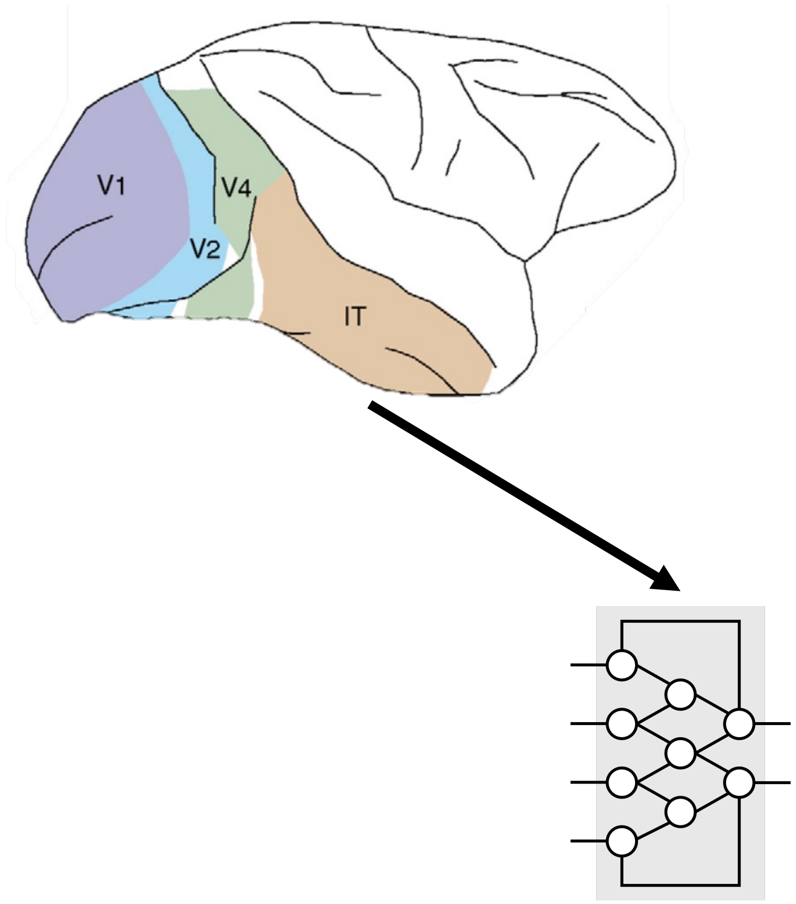
Achieving invariances along visual ventral stream

- Evolving representations achieved more separable object manifolds (Dicarlo & Cox, 2007).



Our question

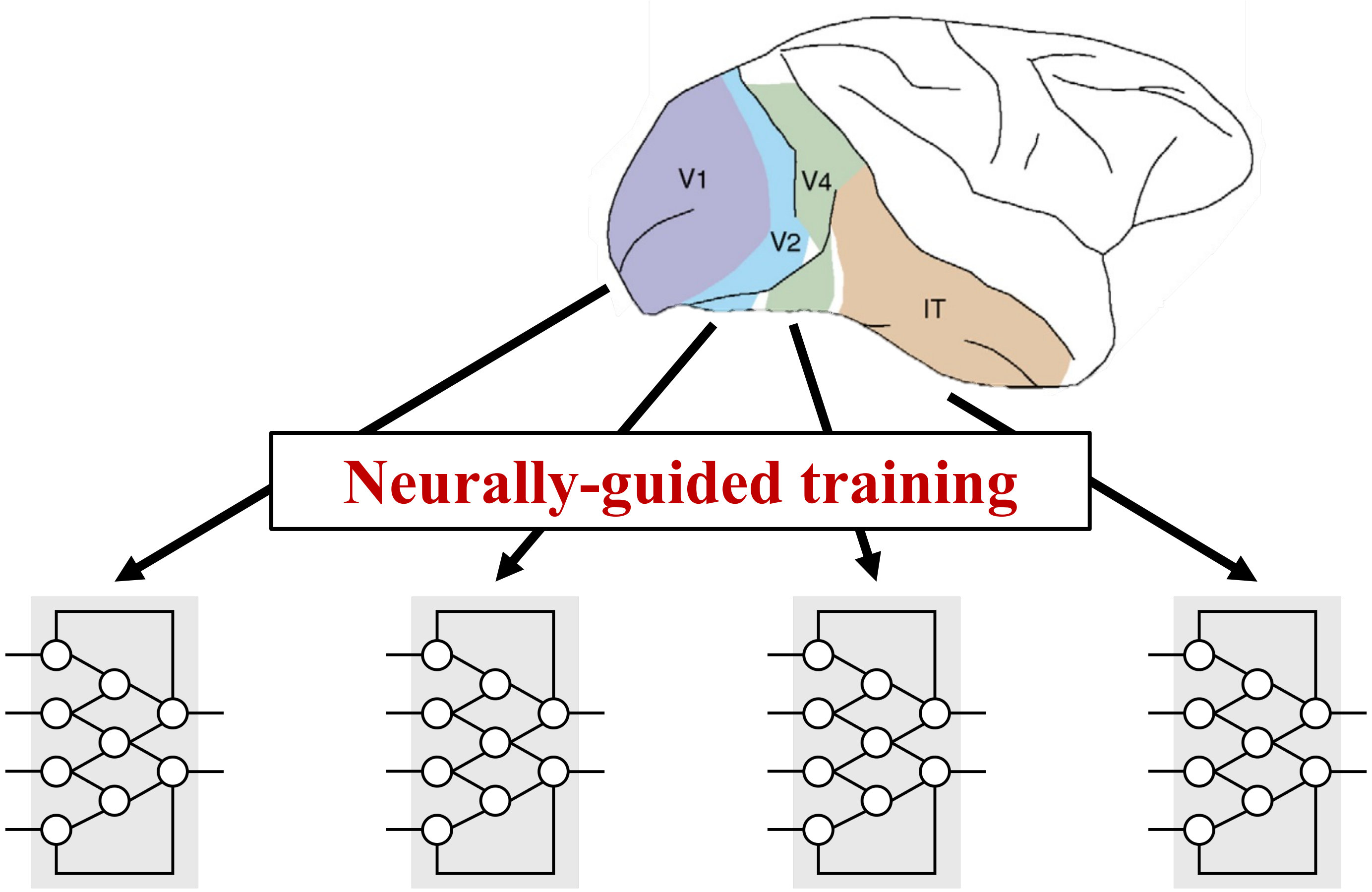
1. Does training guided by human ventral cortex activity improve neural network robustness?



Our question

1. Does training guided by human ventral cortex activity improve neural network robustness?

2. Does such improvement increase as we ascend the ventral visual cortex?

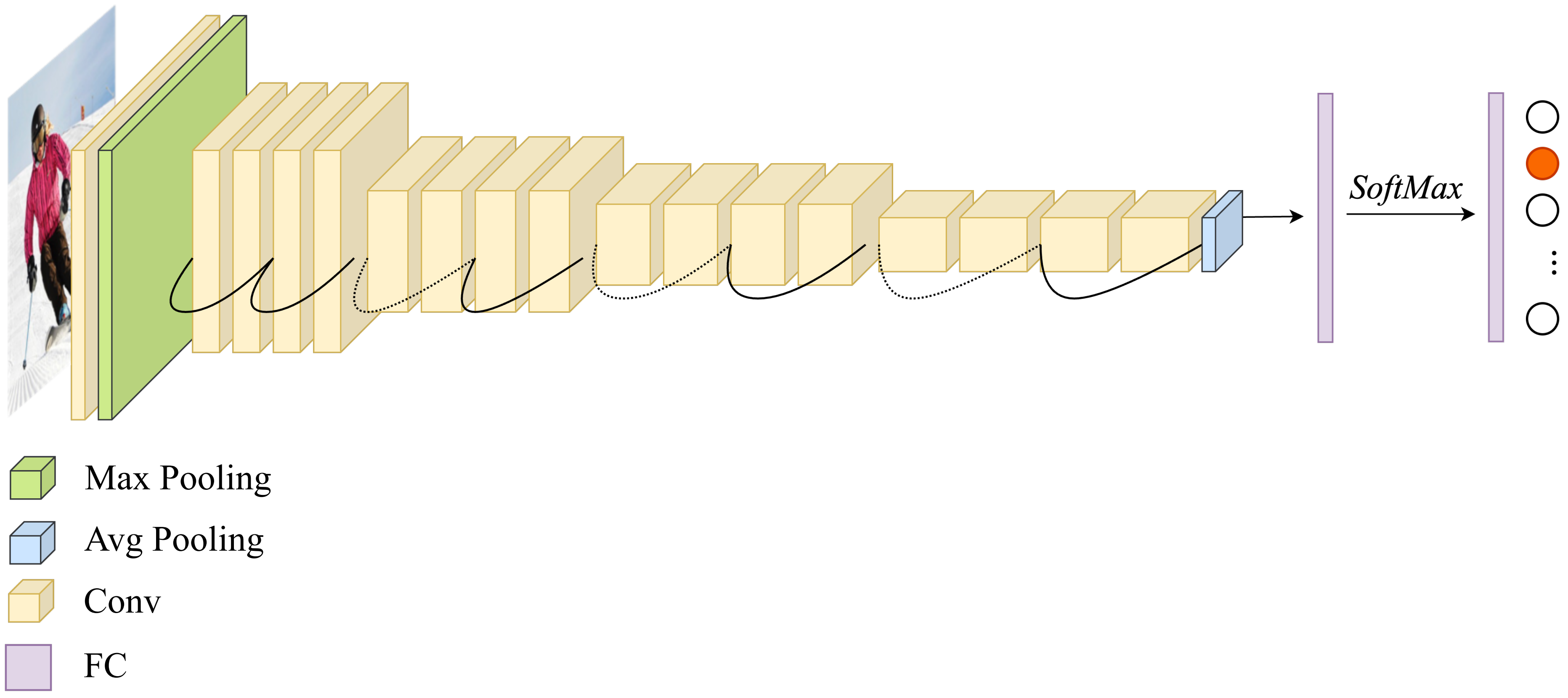


Method

Neurally-guided training

- DNN visual task training:

$$L = L_{task}$$

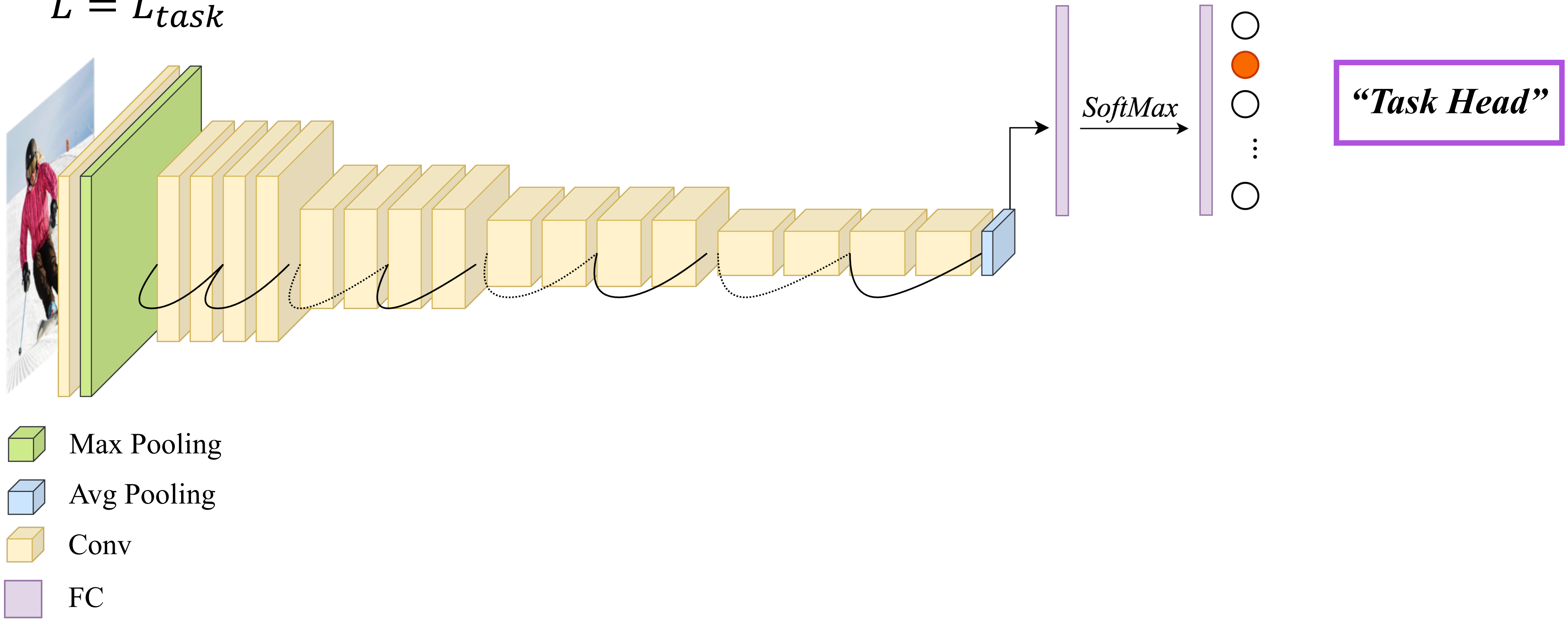


Method

Neurally-guided training

- DNN visual task training:

$$L = L_{task}$$

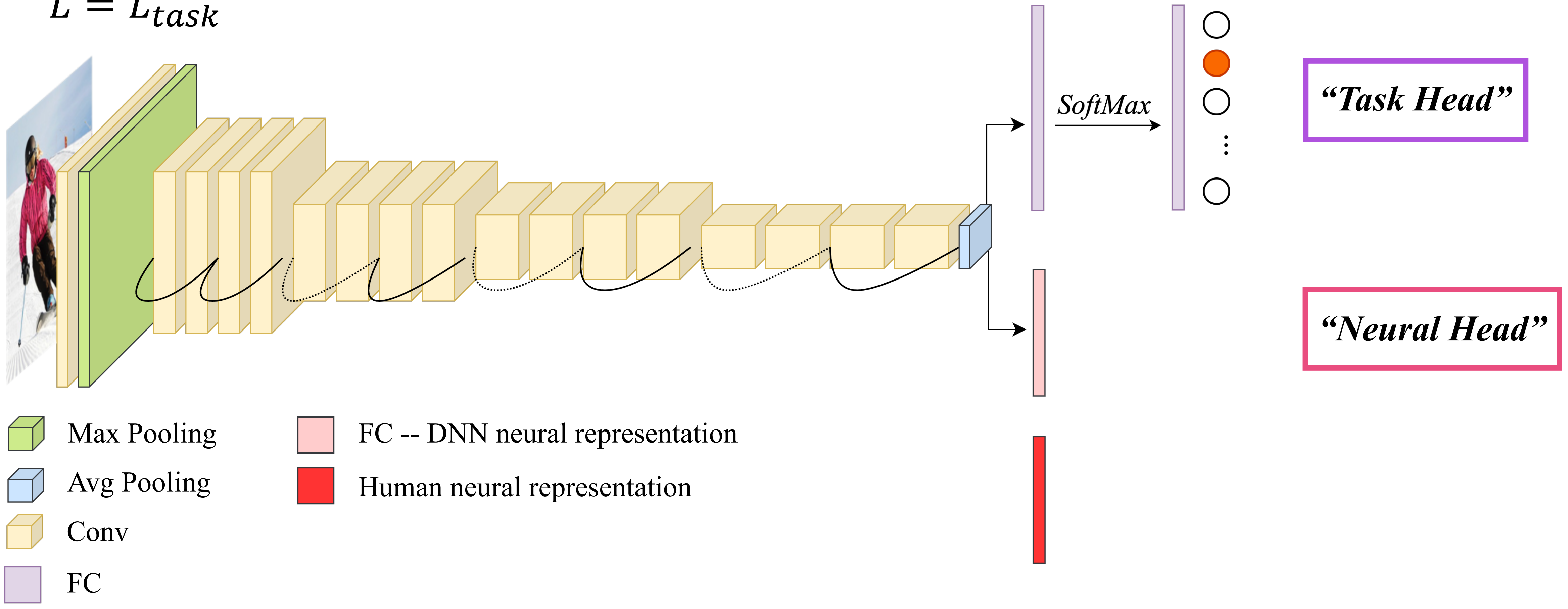


Method

Neurally-guided training

- DNN visual task training with Neural Guidance:

$$L = L_{task}$$

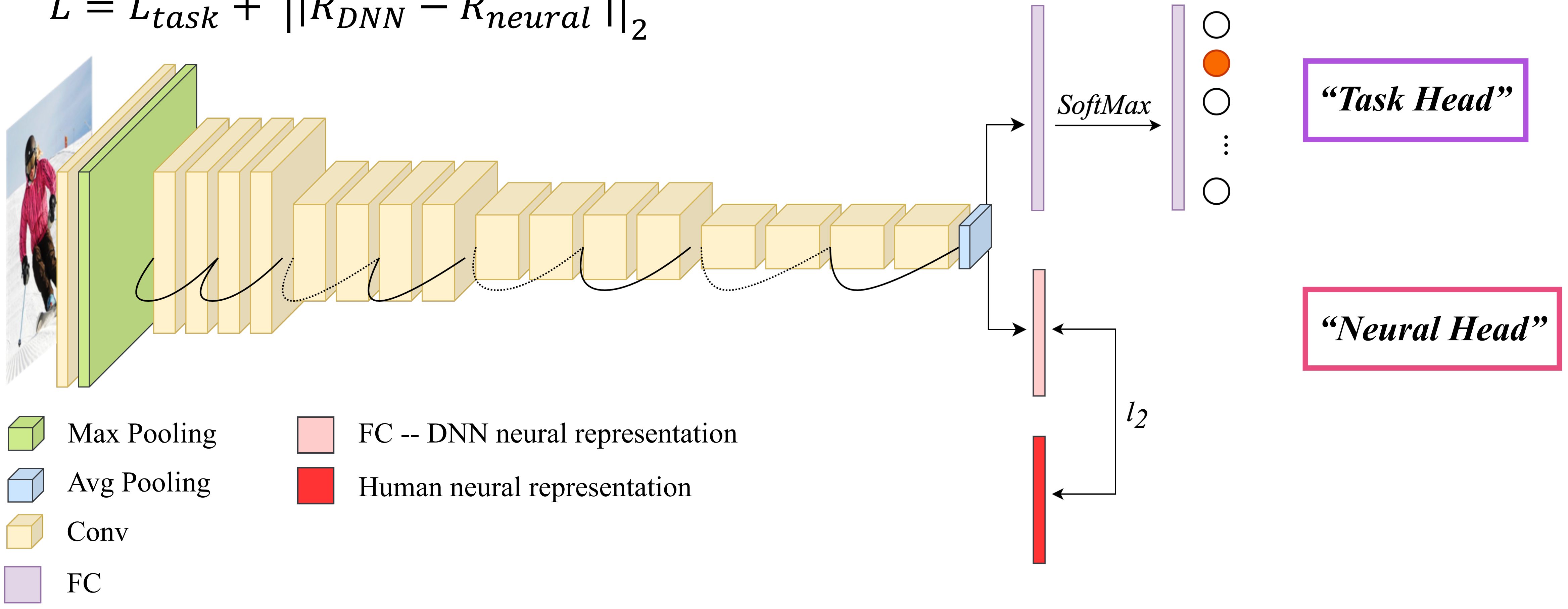


Method

Neurally-guided training

- DNN visual task training with Neural Guidance:

$$L = L_{task} + \left\| R_{DNN} - R_{neural} \right\|_2$$

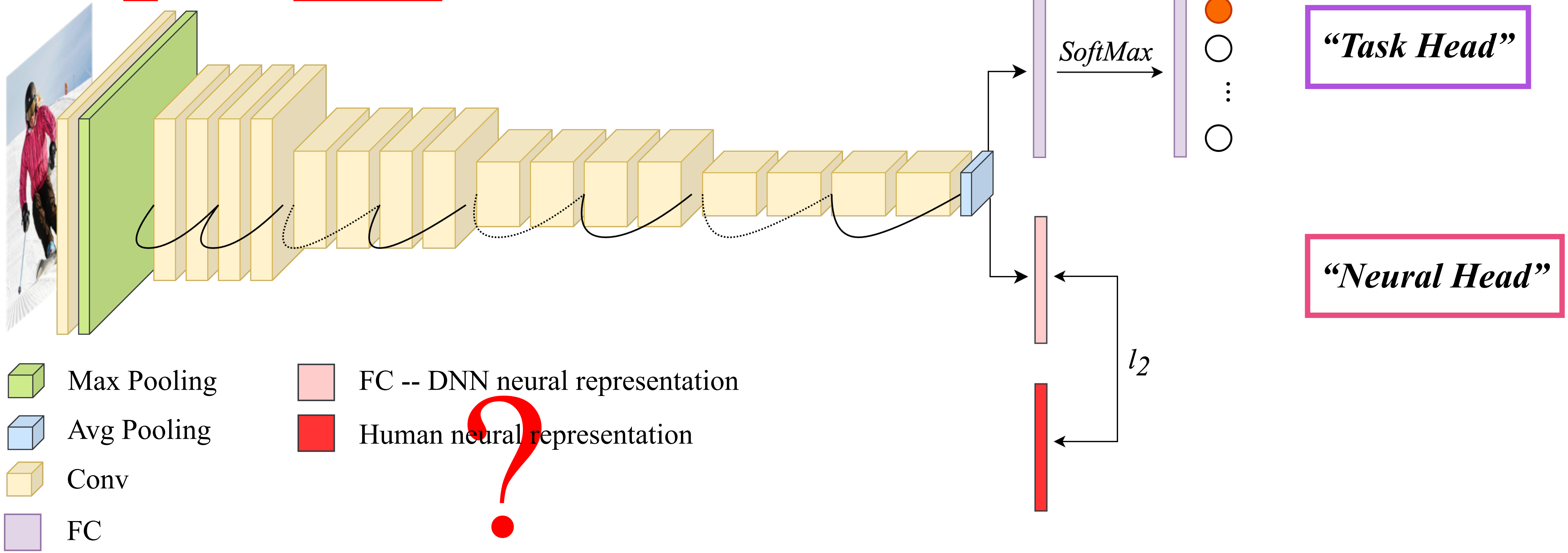


Method

Neurally-guided training

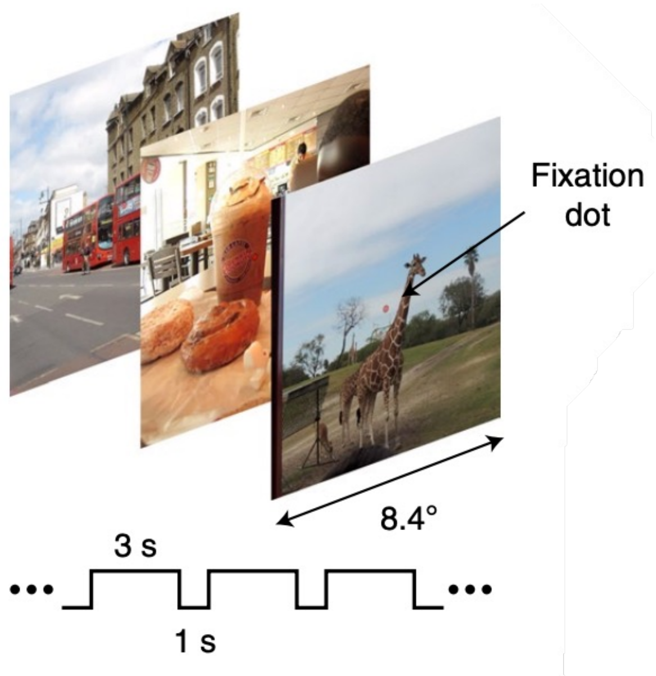
- DNN visual task training with Neural Guidance:

$$L = \boxed{\alpha} L_{task} + \boxed{(1-\alpha)} ||R_{DNN} - R_{neural}||_2$$



Method
Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.

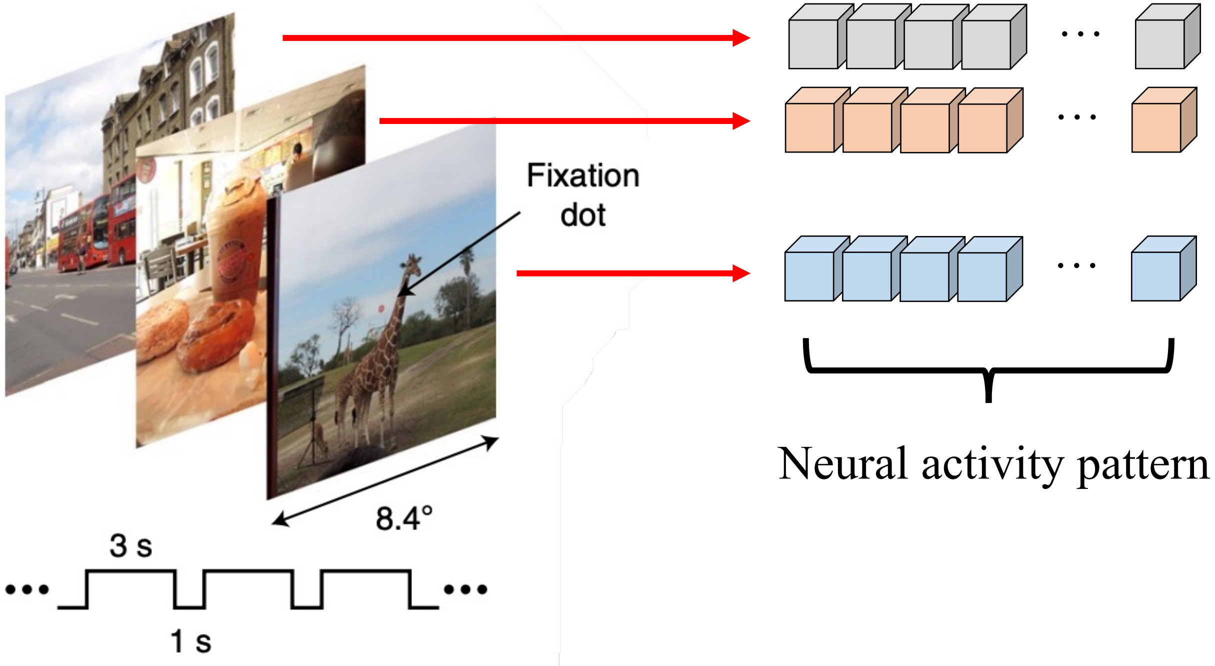


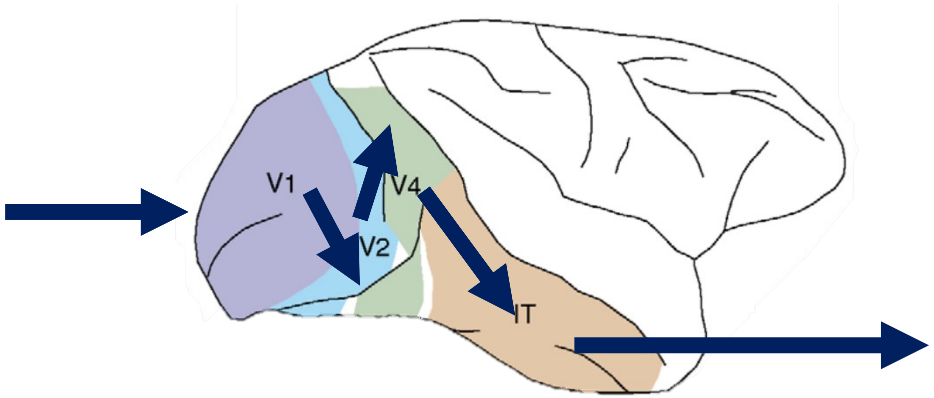
(NSD, Allen et al., 2022)

Method

Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.

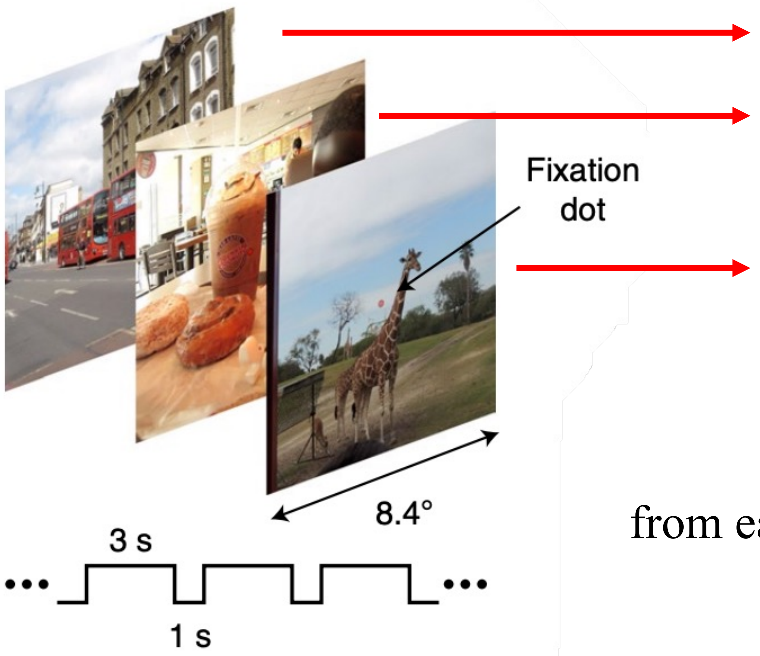


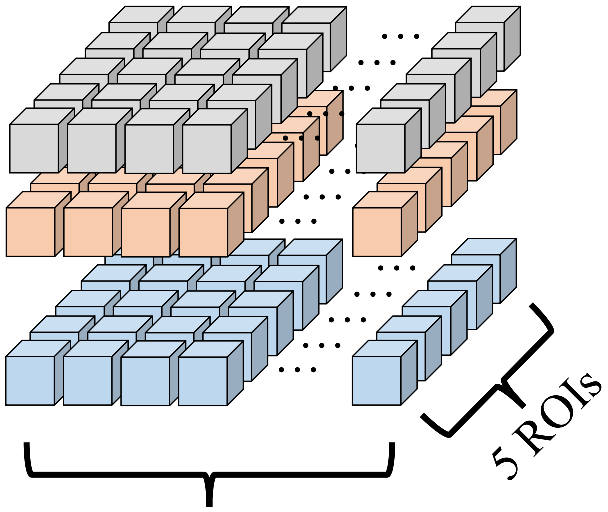


(NSD, Allen et al., 2022)

Method
Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral Regions of Interest (**ROIs**) were used





Neural activity pattern
from each ROI of the ventral hierarchy

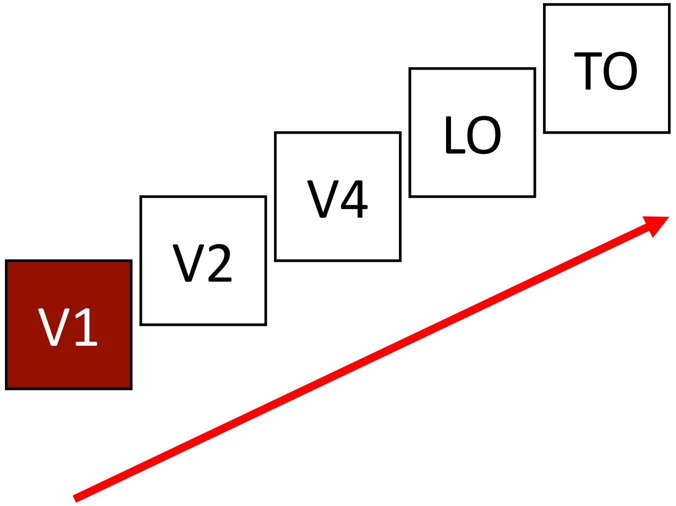
Subject 1 ROIs

(NSD, Allen et al., 2022)

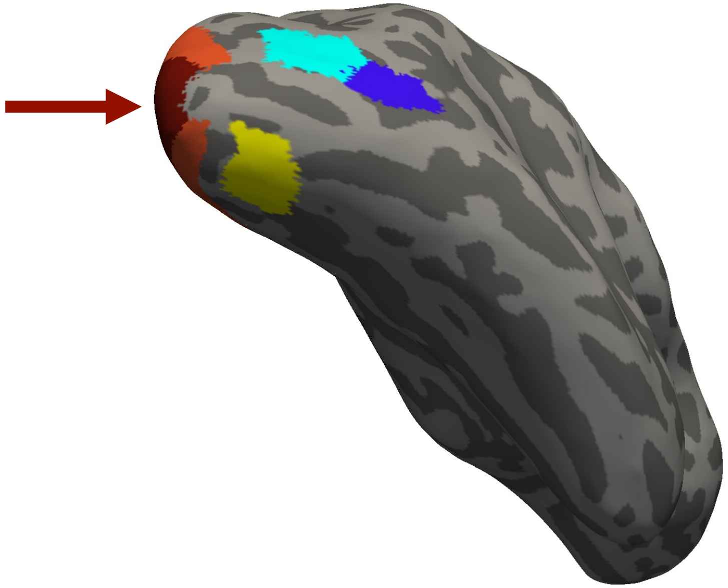
Method

Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used:



Ventral Visual Stream
Hierarchy Order

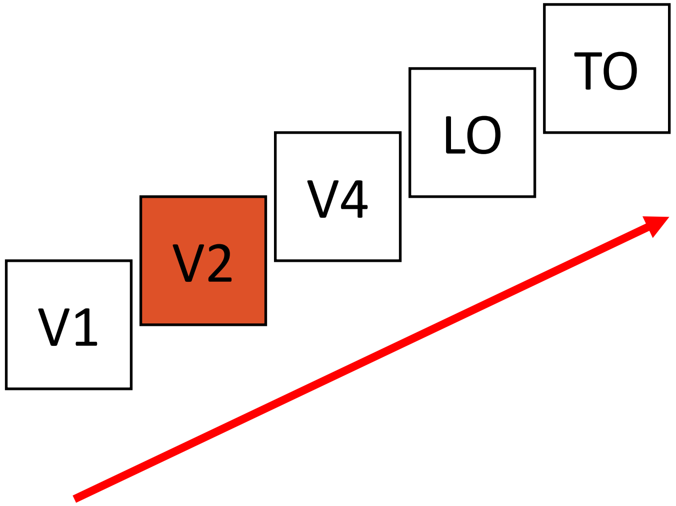


Subject 1 ROIs

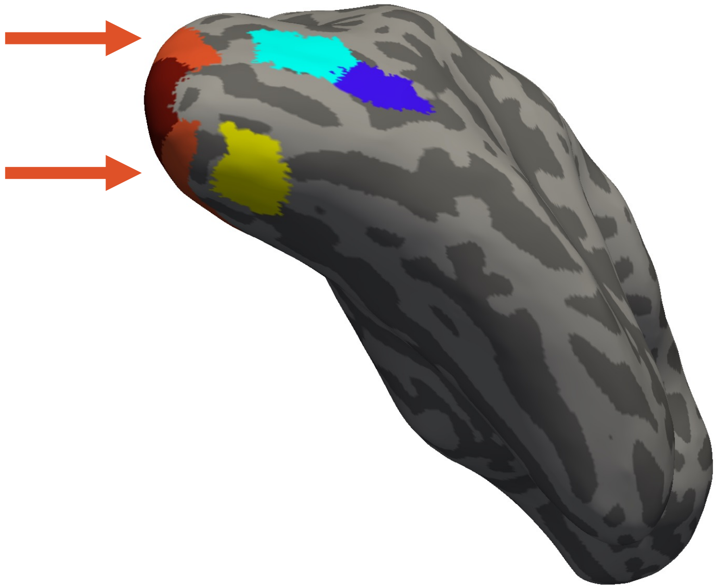
Method

Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used:



Ventral Visual Stream Hierarchy Order

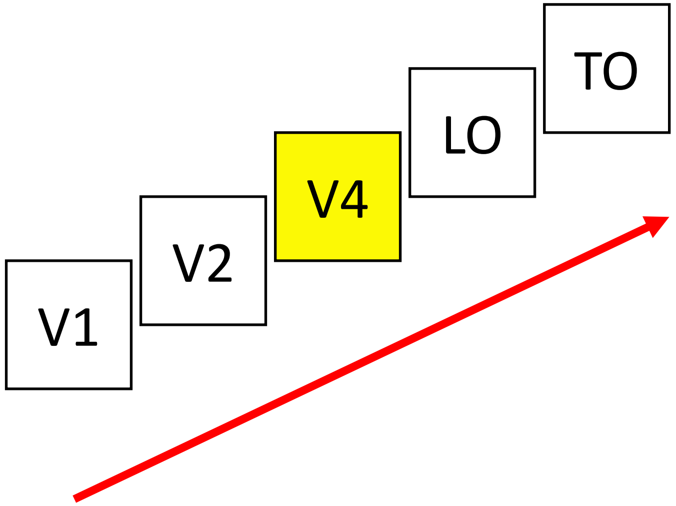


Subject 1 ROIs

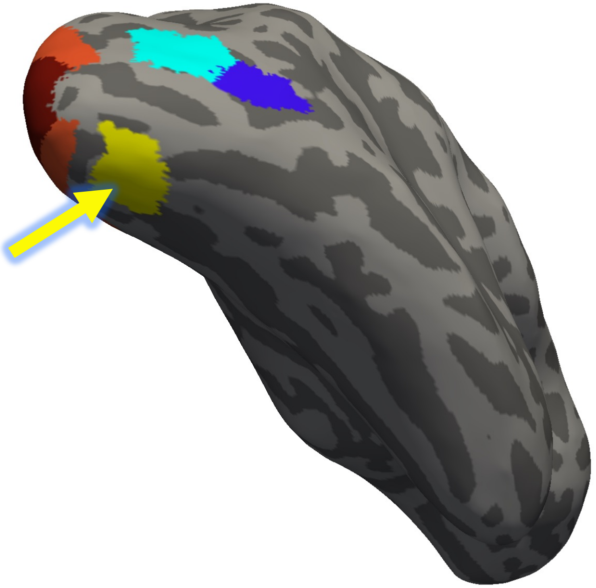
Method

Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used:



Ventral Visual Stream Hierarchy Order

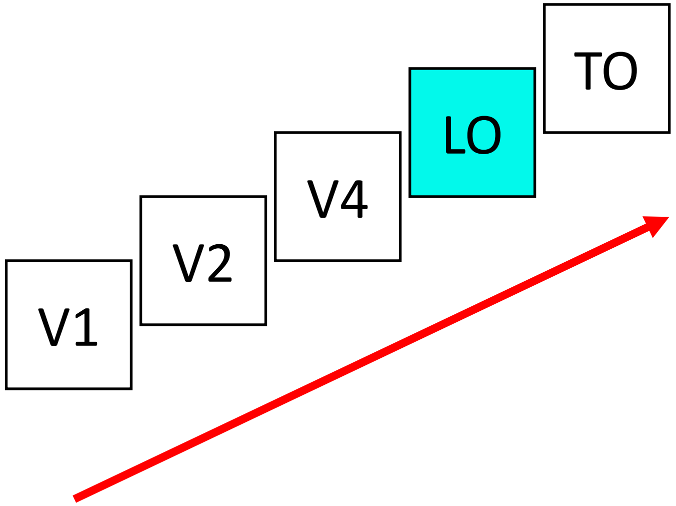


Subject 1 ROIs

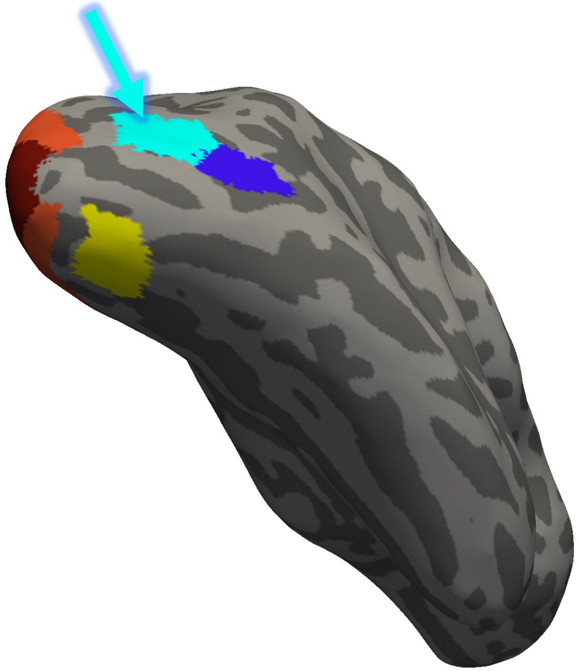
Method

Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used:



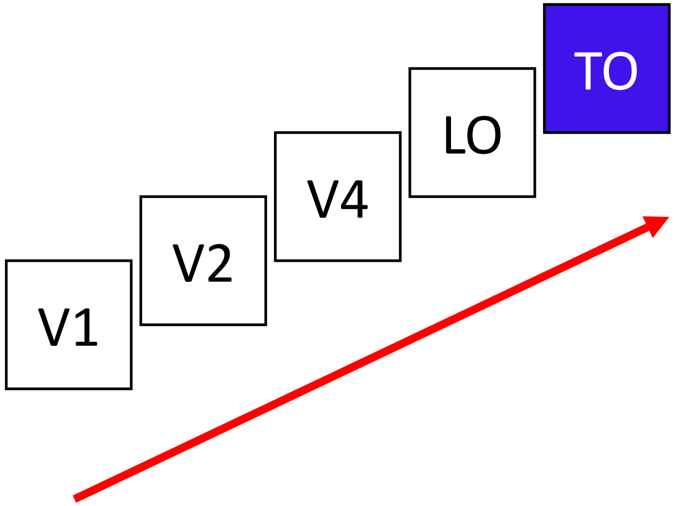
Ventral Visual Stream
Hierarchy Order



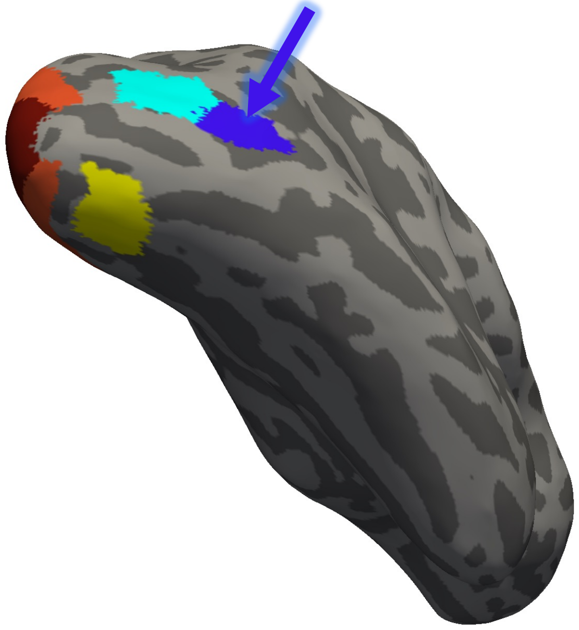
Subject 1 ROIs

Method
Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used:



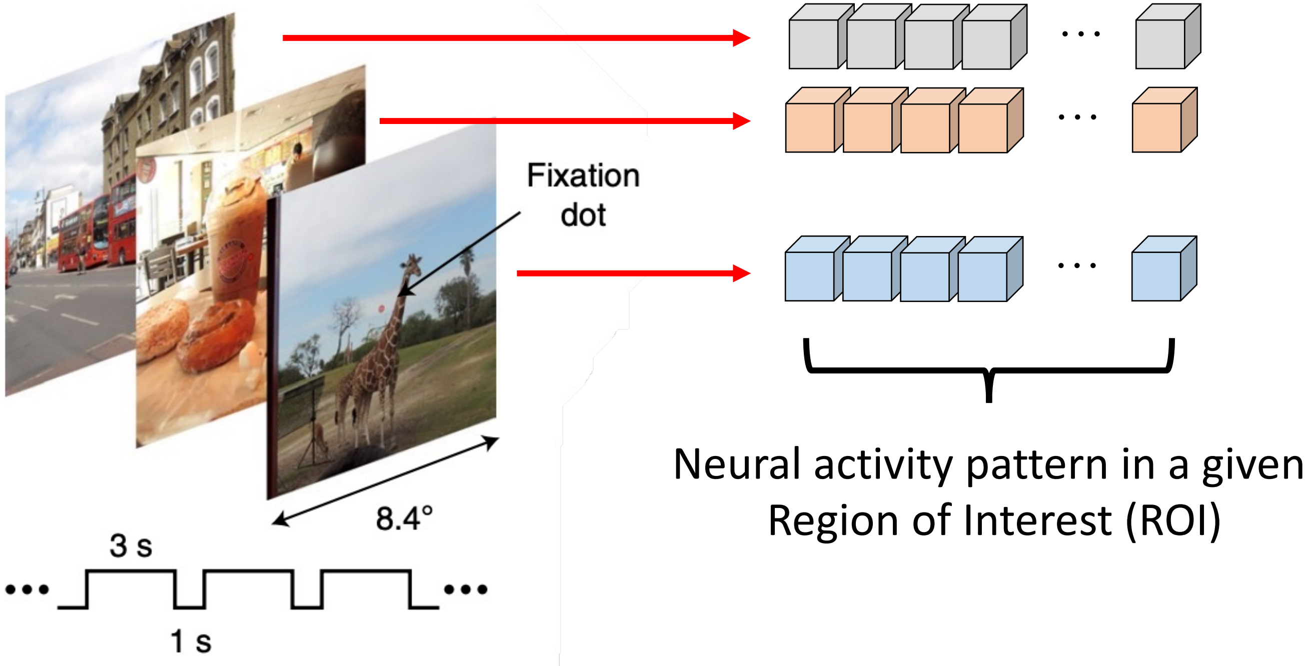
Ventral Visual Stream Hierarchy Order

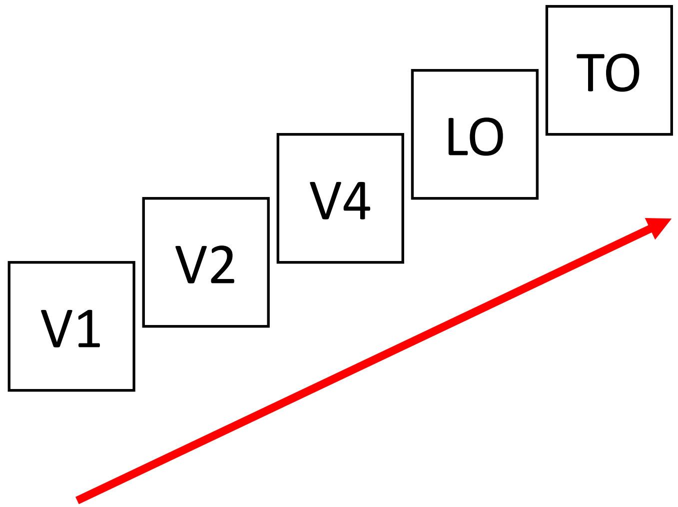


Subject 1 ROIs

Method
Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used:



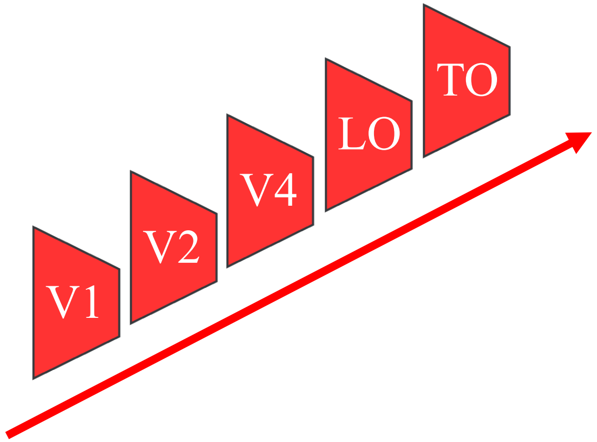


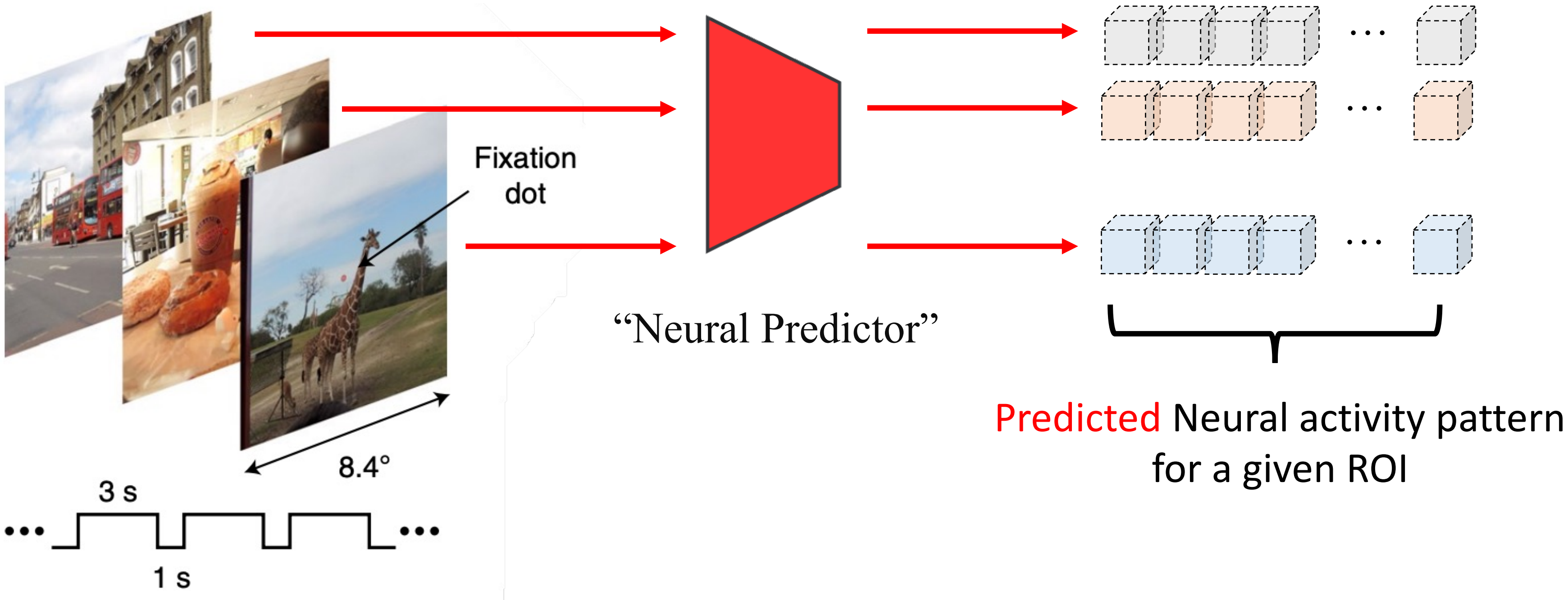
(NSD, Allen et al., 2022)

Method

Neural data

- Each human subject viewed ~30,000 images (~9,000 unique).
- Brain activities were recorded with 7T fMRI.
- 5 bilateral ROIs were used.





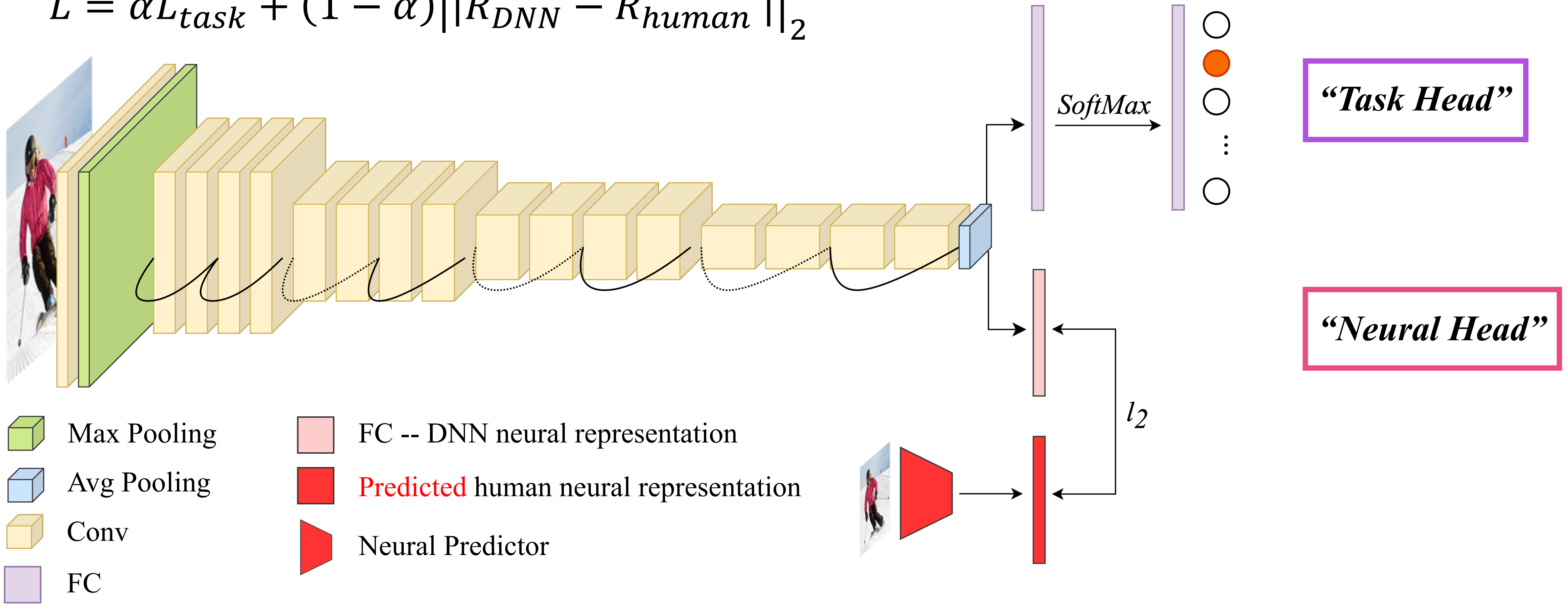
(NSD, Allen et al., 2022)

Method

Neurally-guided training

- DNN visual task training with Neural Guidance:

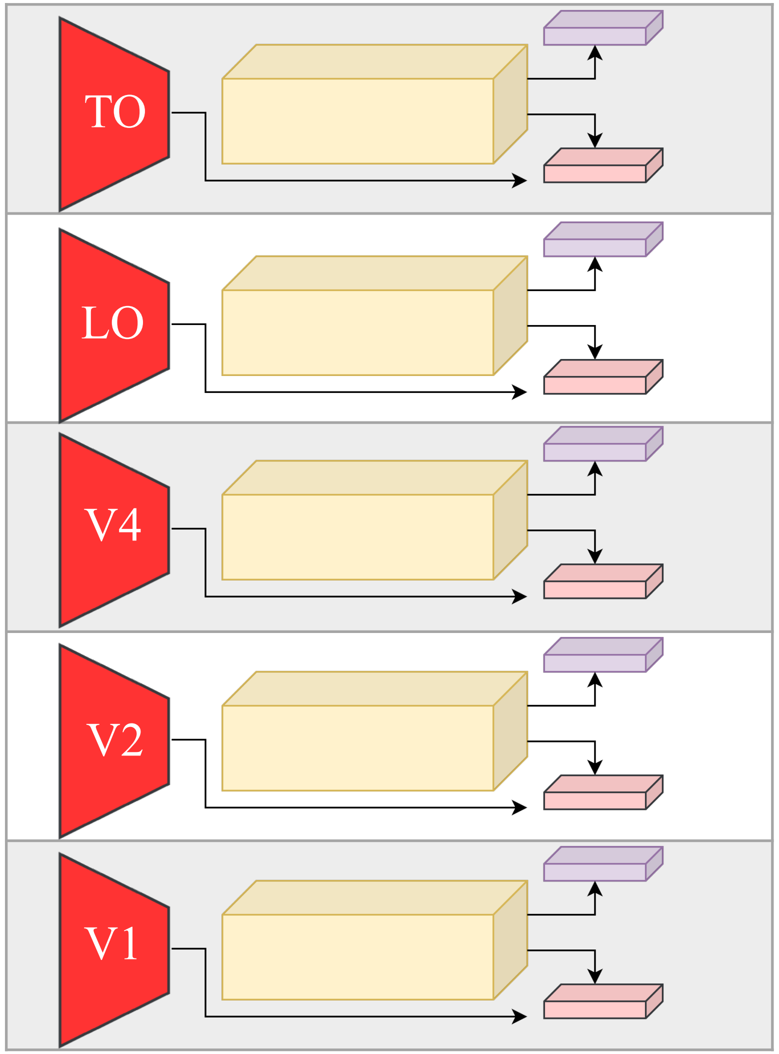
$$L = \alpha L_{task} + (1-\alpha)||R_{DNN} - R_{human}||_2$$



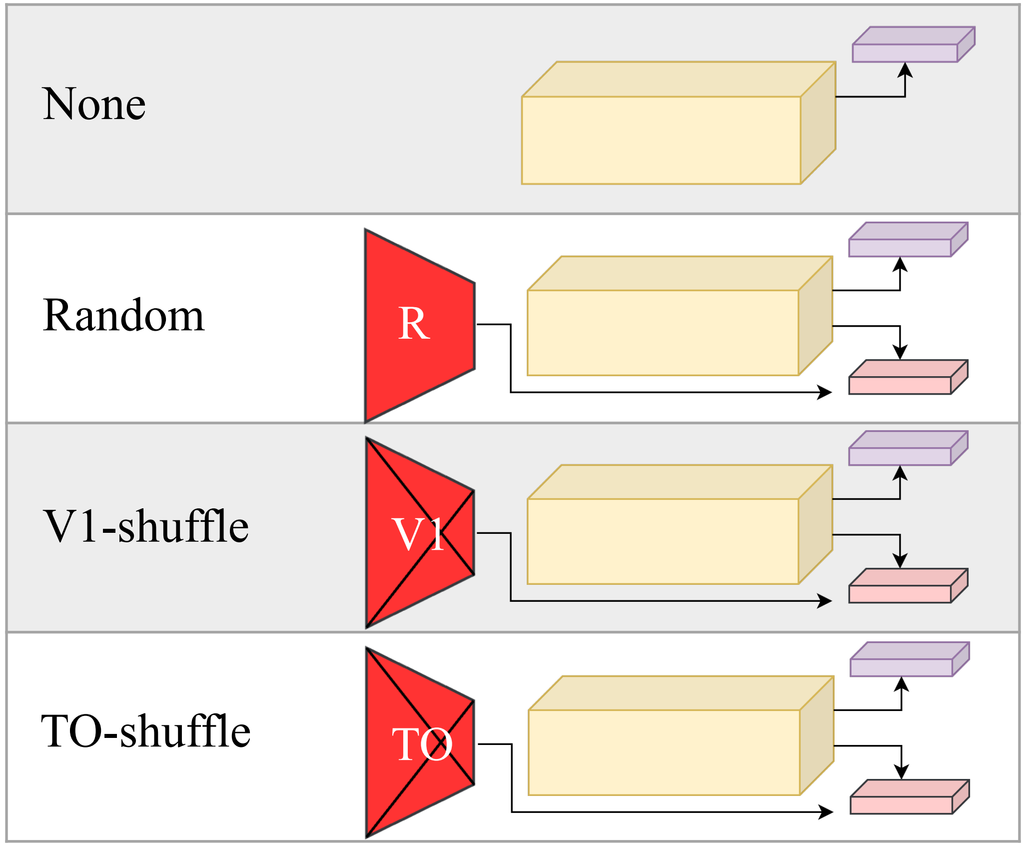
Method

Summary of models

- 5 models with neural-guided training



- 4 baseline models for comparison

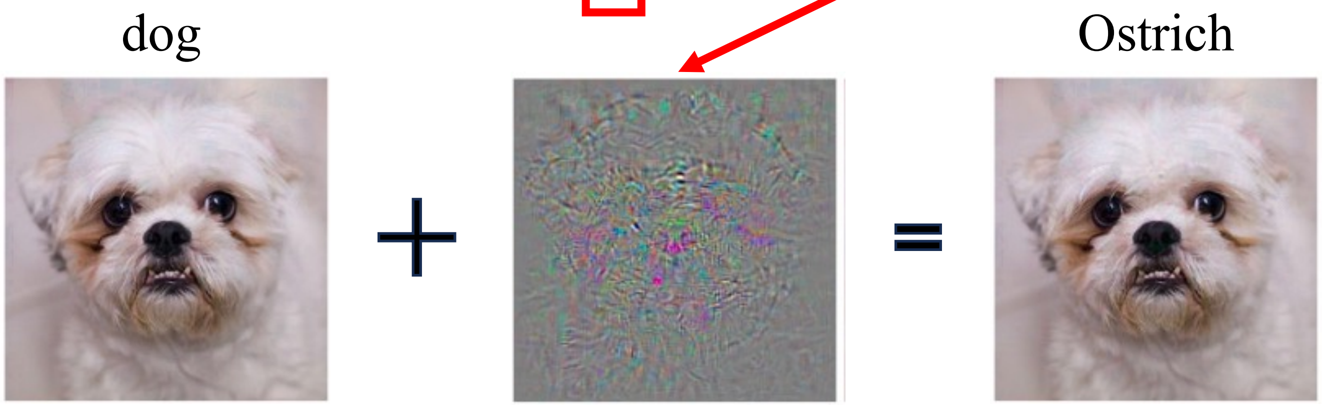


Robustness of Neurally-guided Models

Evaluation

- **l_p-based adversarial attack:**

$$\max_{||\tau||_p < \boxed{\epsilon}} l(f_\theta(x+\tau), y)$$



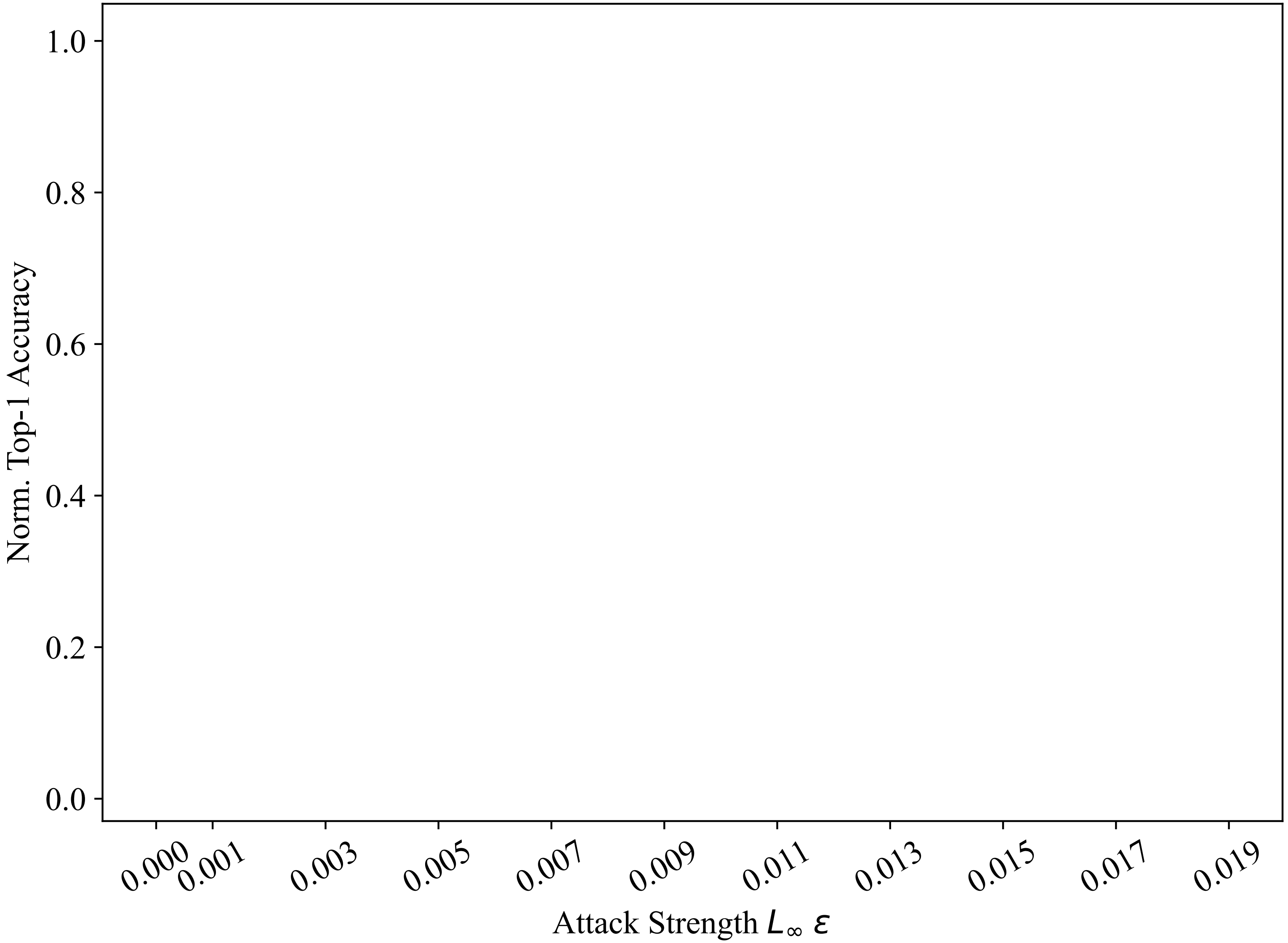
(Szegedy et al., 2014)

We check:
1. Does neural guidance improve neural network robustness?
2. Is such improvement hierarchical?

Results

l_∞-based PGD adversarial attack

Task: Image Classification
Dataset: ImageNet (Deng et al., 2009)

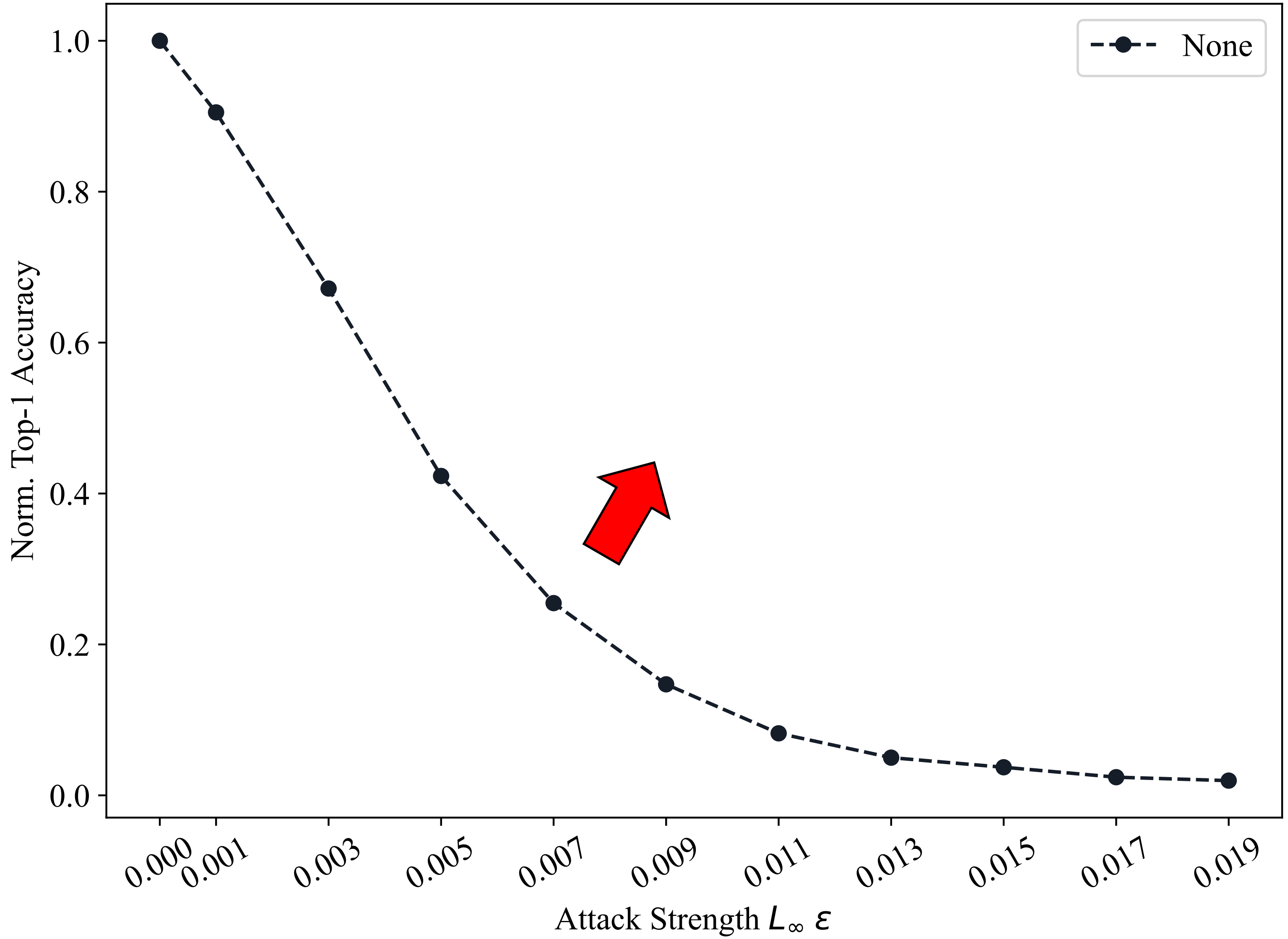


“None”

Results

l_∞-based PGD adversarial attack

Task: Image Classification
Dataset: ImageNet (Deng et al., 2009)



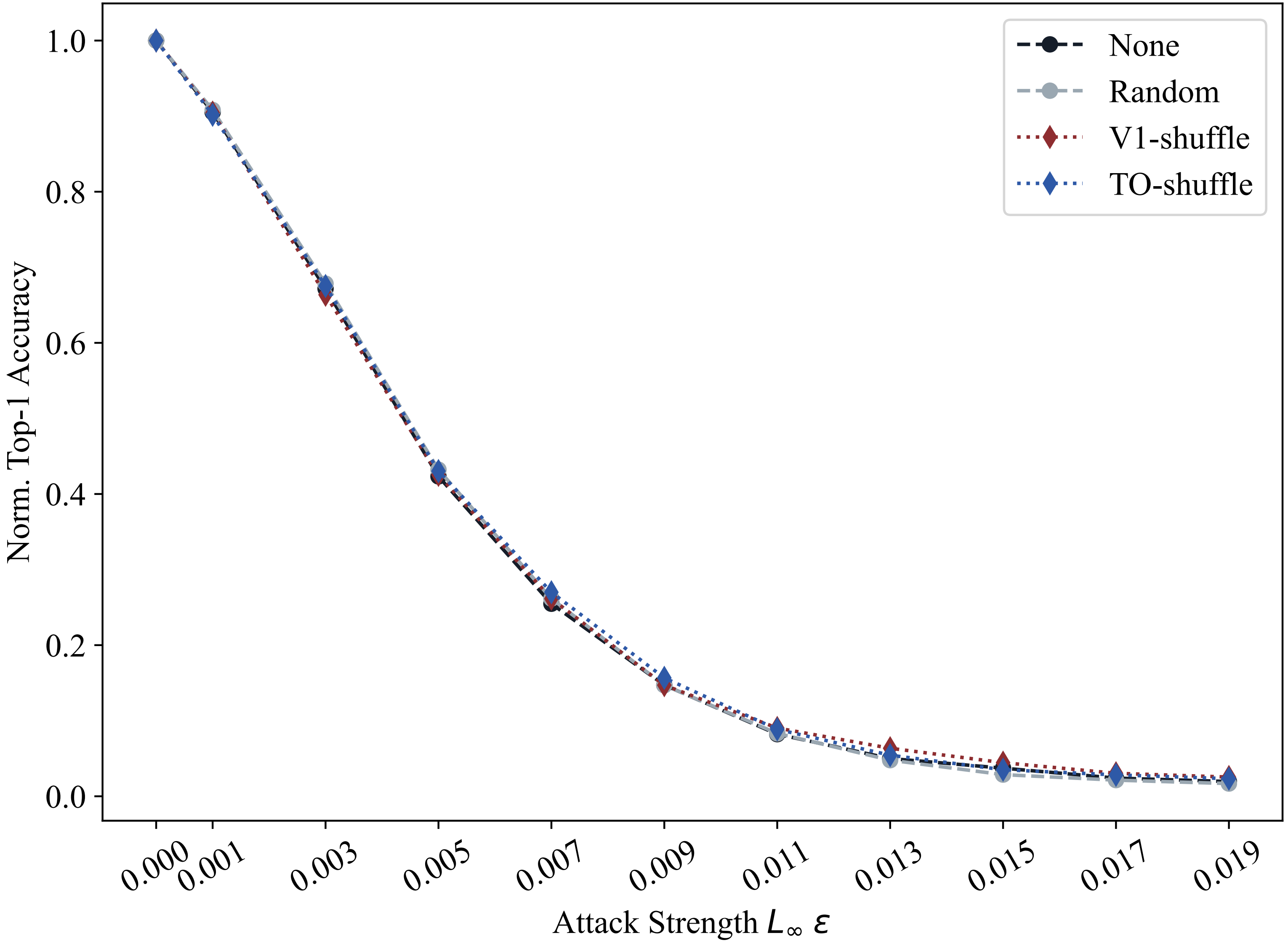
"None"

Results

l_∞-based PGD adversarial attack

Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)



"None" "Random"

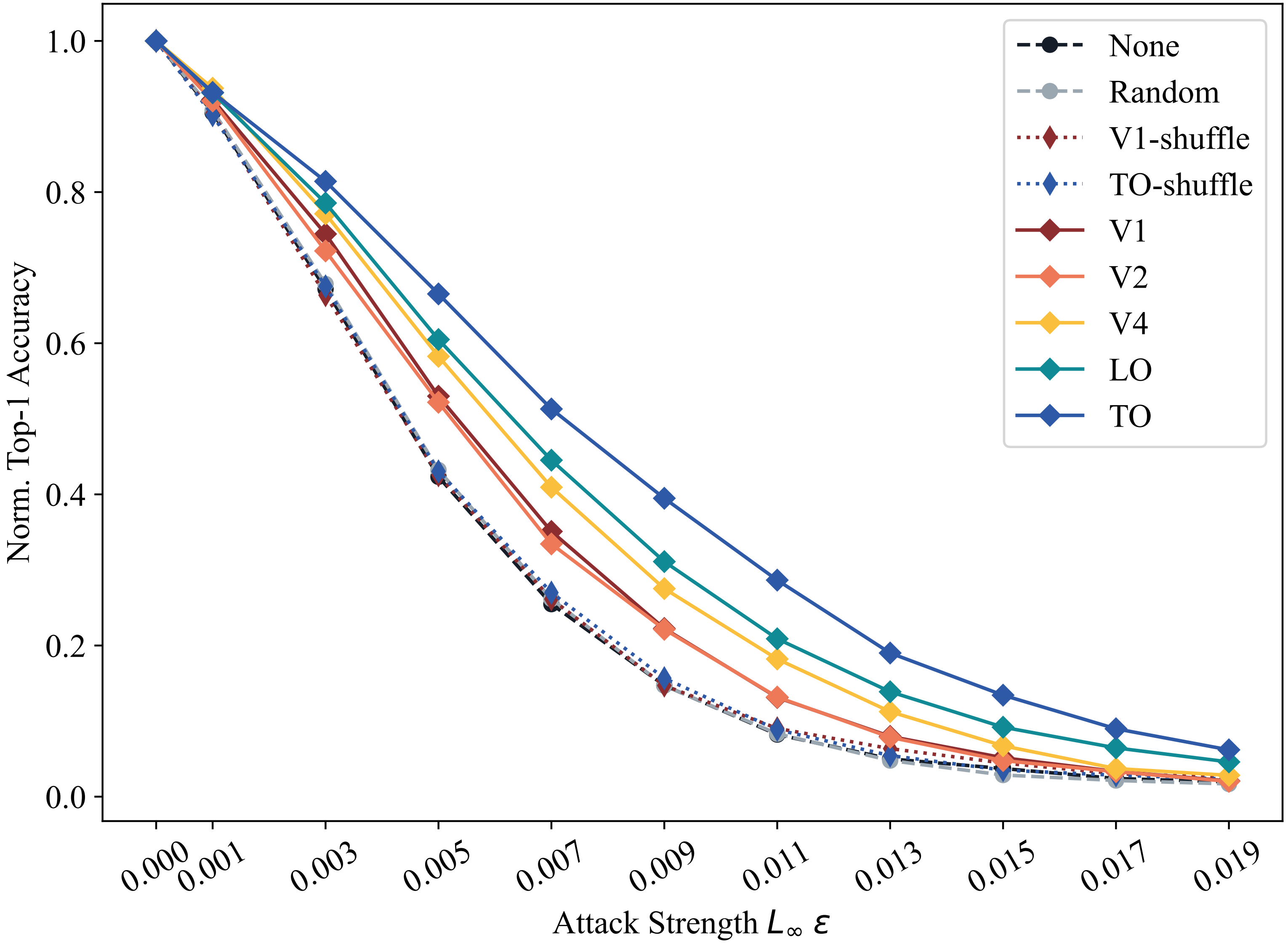
"V1-shuffle" "TO-shuffle"

Results

l_∞-based PGD adversarial attack

Task: Image Classification

Dataset: ImageNet (Deng et al., 2009)



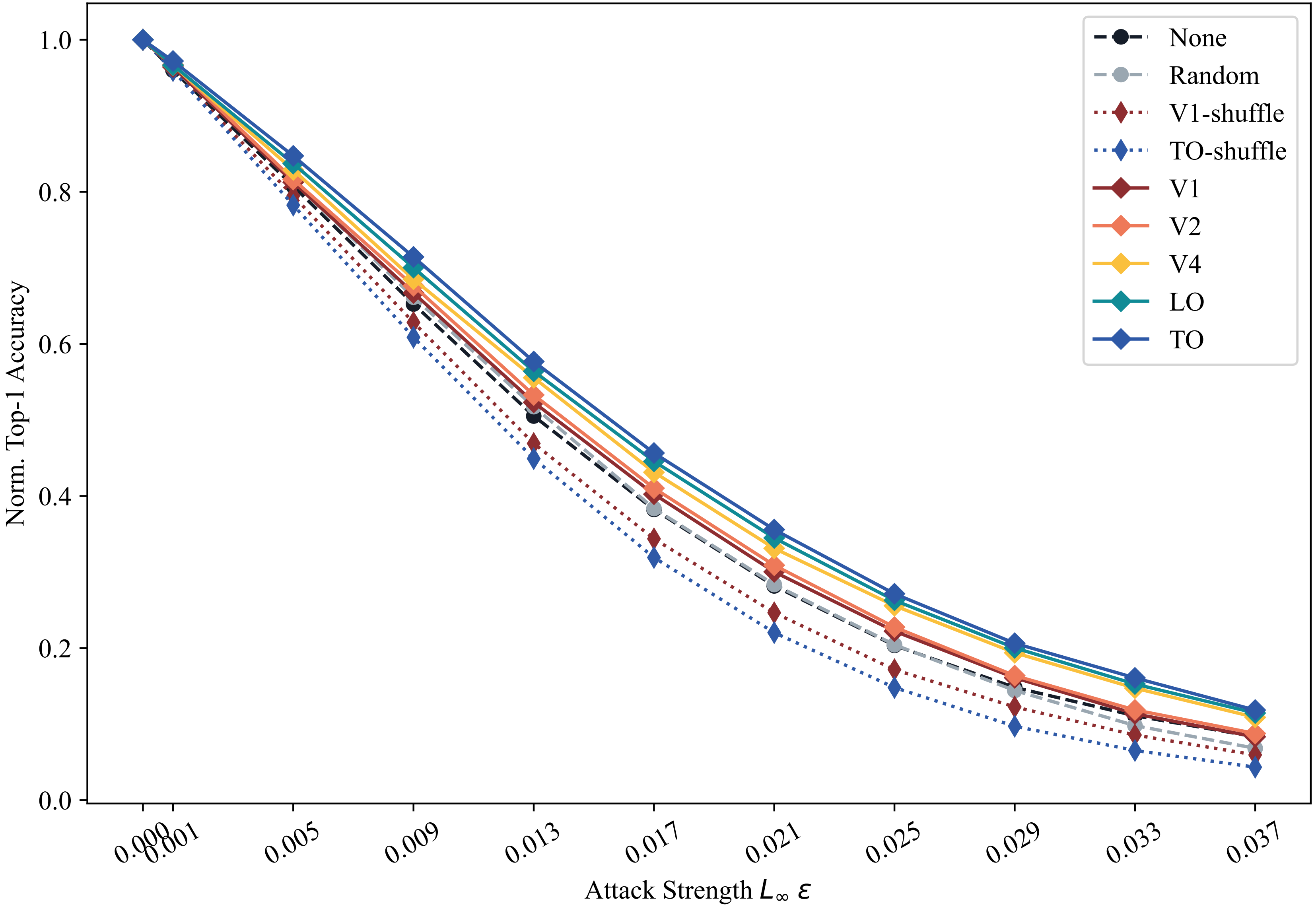
- **Neural guidance improves robustness** (max: 22% accuracy increase)
- **There exists a hierarchy of improvement's magnitude**
- **Similar results have been observed with:**
  - L_∞ FGSM
  - Auto-Attack (APGD-CE, APGD-T, FAB square)
  - L_2 FGM
  - L_2 Deepfool

Results

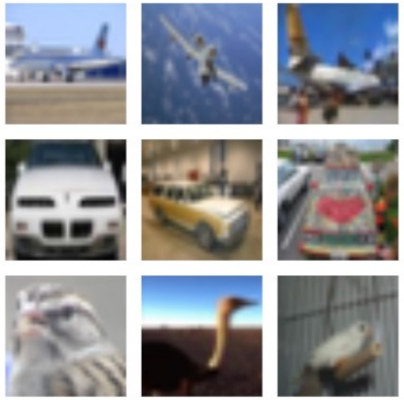
l_∞-based PGD adversarial attack

Task: Image Classification

Dataset: CIFAR-100 (Krizhevsky, 2009)



- **Neural guidance improves robustness** (max: 12% accuracy increase)
- **There exists a hierarchy of improvement's magnitude**

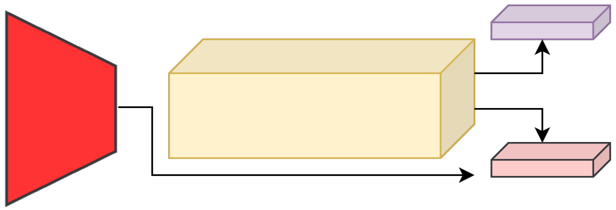


Results

l_∞-based PGD adversarial attack

Task: Image Captioning

Dataset: MSCOCO (Lin et al, 2014)

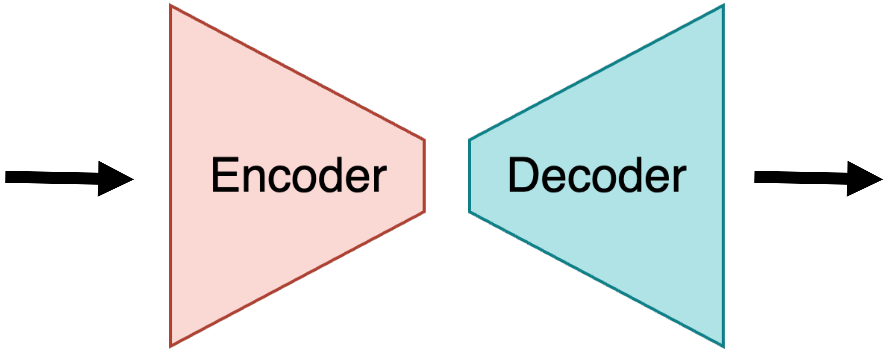


Results

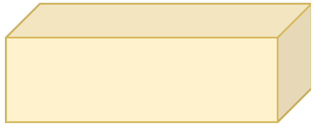
l_∞-based PGD adversarial attack

Task: Image Captioning
Dataset: MSCOCO (Lin et al, 2014)

- Do representations from neurally-guided DNNs benefit other visual tasks beyond basic classification?



"Happy dog sitting in the bed of a pickup truck."



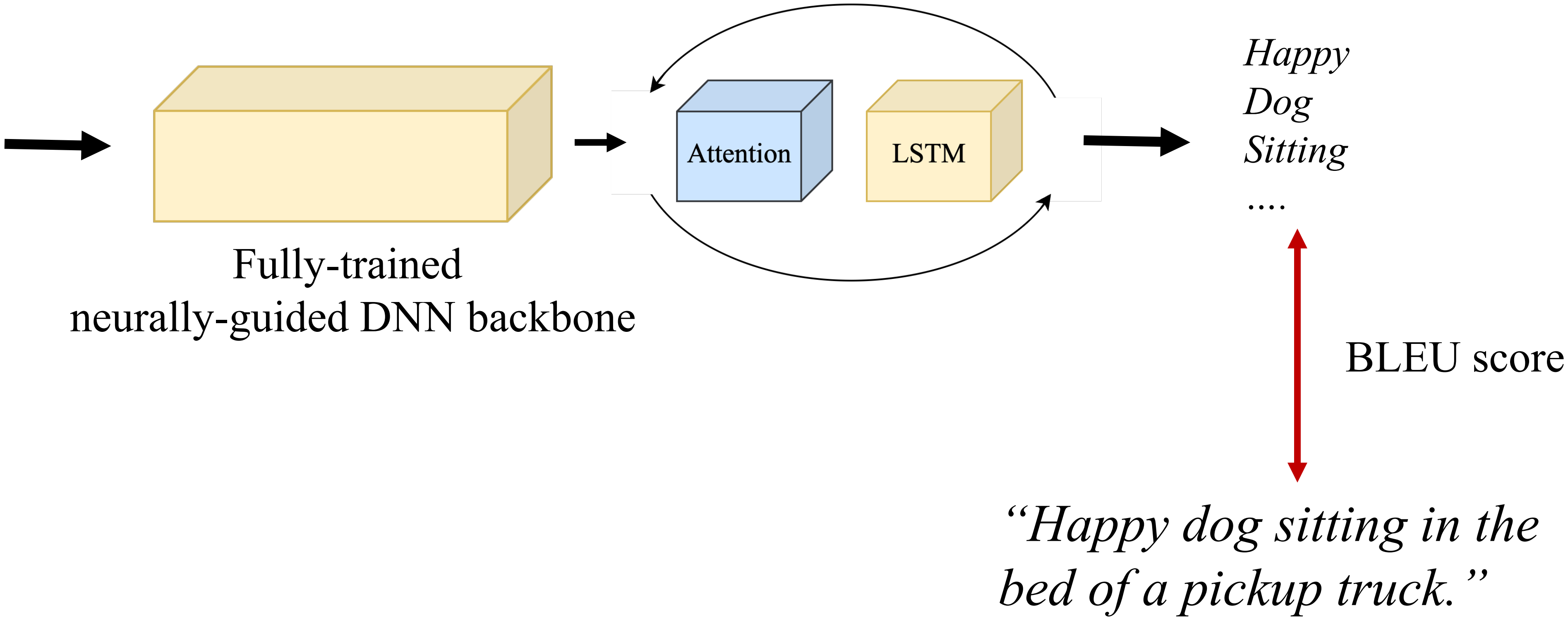
("Show, Attend, &Tell", Xu et al., 2015)

Results

l_∞-based PGD adversarial attack

Task: Image Captioning
Dataset: MSCOCO (Lin et al, 2014)

- Do representations from neurally-guided DNNs benefit other visual tasks beyond basic classification?

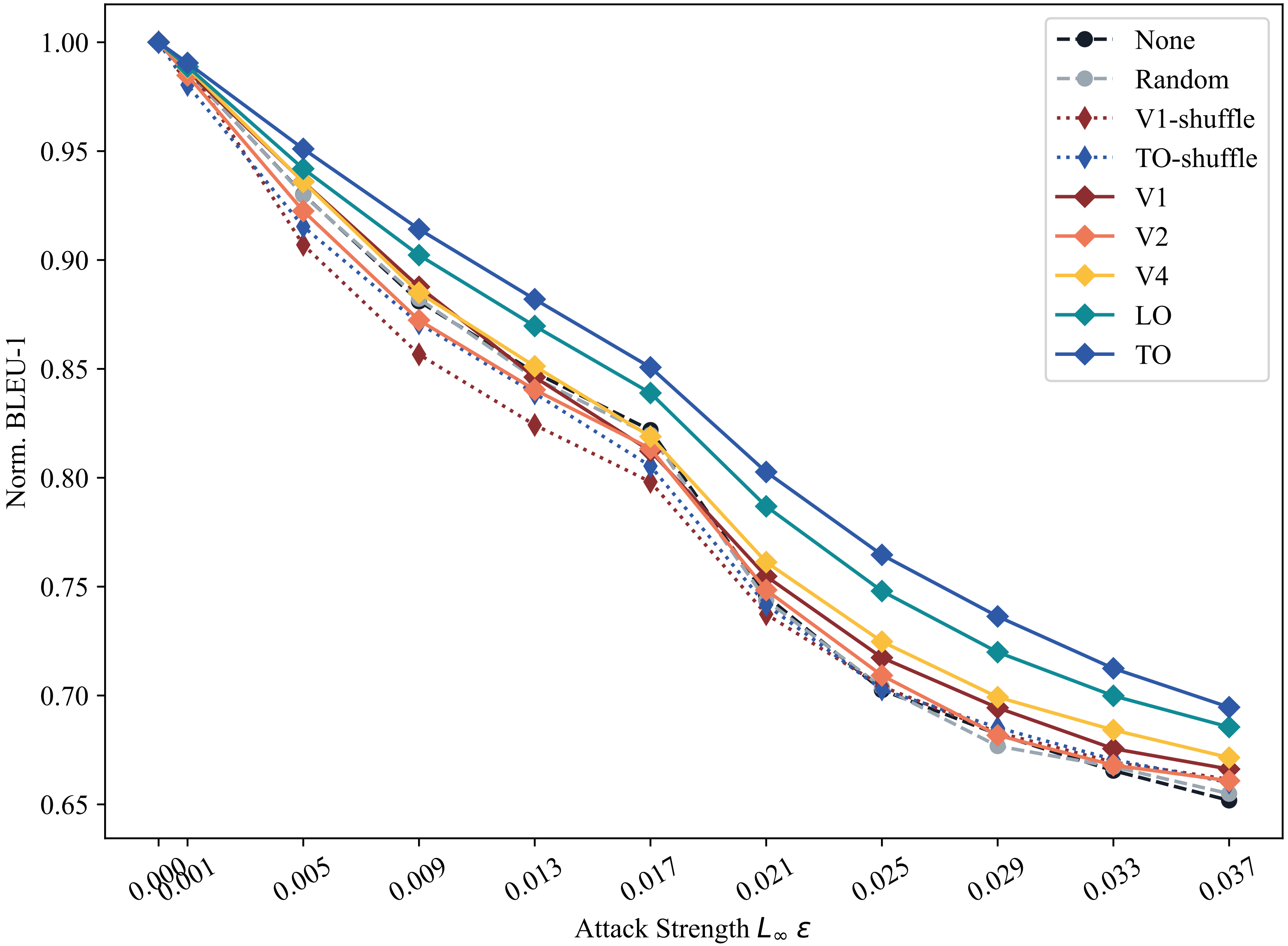


("Show, Attend, &Tell", Xu et al., 2015)

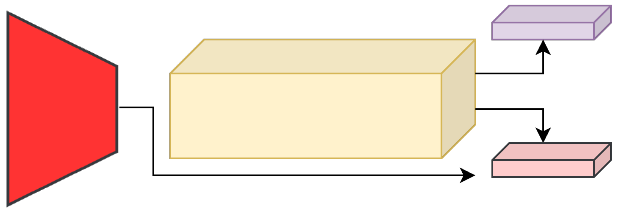
Results

l_∞-based PGD adversarial attack

Task: Image Captioning
Dataset: MSCOCO (Lin et al., 2014)

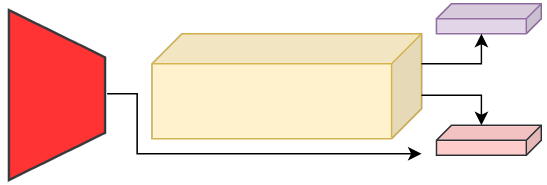


- **Neural regularization improves robustness** (max: 0.03 BLEU-1 increase)
- **There exists a hierarchy of improvement's magnitude**

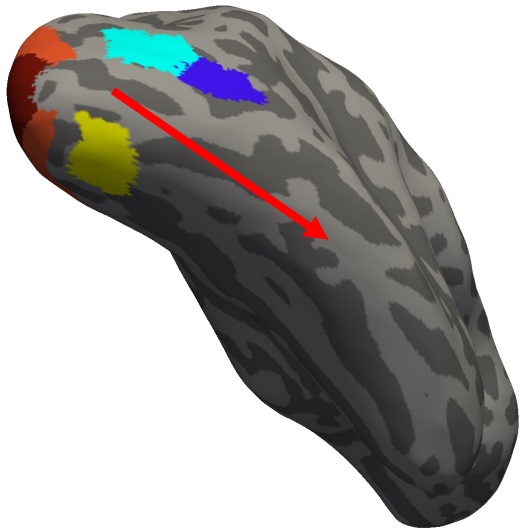


Neural-guidance
→ Robust feature extractor

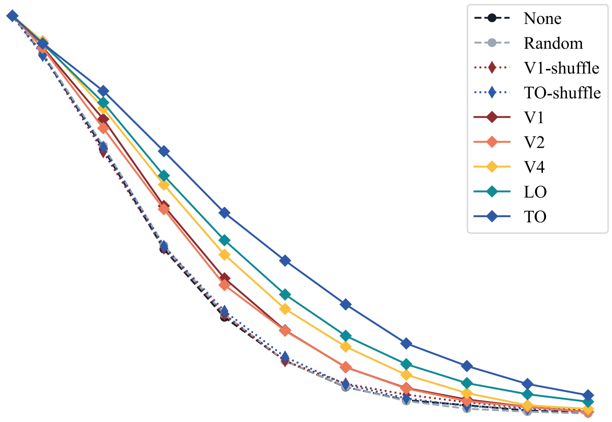
Conclusion & Discussion



- We found hierarchical improvements in DNN robustness across:
  - Datasets (ImageNet, CIFAR-100, MSCOCO)
  - Tasks (Classification, Captioning)
  - Attacks (L_∞ PGD, L_∞ FGSM, Autoattack, L_2 FGM, L_2 Deepfool)
- Implications:
  - Evolving representation space along ventral visual stream
  - Learnable and improvable with generic DNN structures
  - Potential for uncovering principles of building human-like representation space and advancing DNN architectural development



- Further analysis
  - Neurally-guided models are more shape-biased
  - Smoother output surface achieved in a different way from conventional solutions.
  - Neurally-guided models experience profound changes in their representation space



Acknowledgement

Linjian Ma

Bo Li

Diane M. Beck

ACCESS

NSF

This work used NCSA Delta GPU through allocation SOC230011 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.