Visual Attention Emerges from Recurrent Sparse Reconstruction

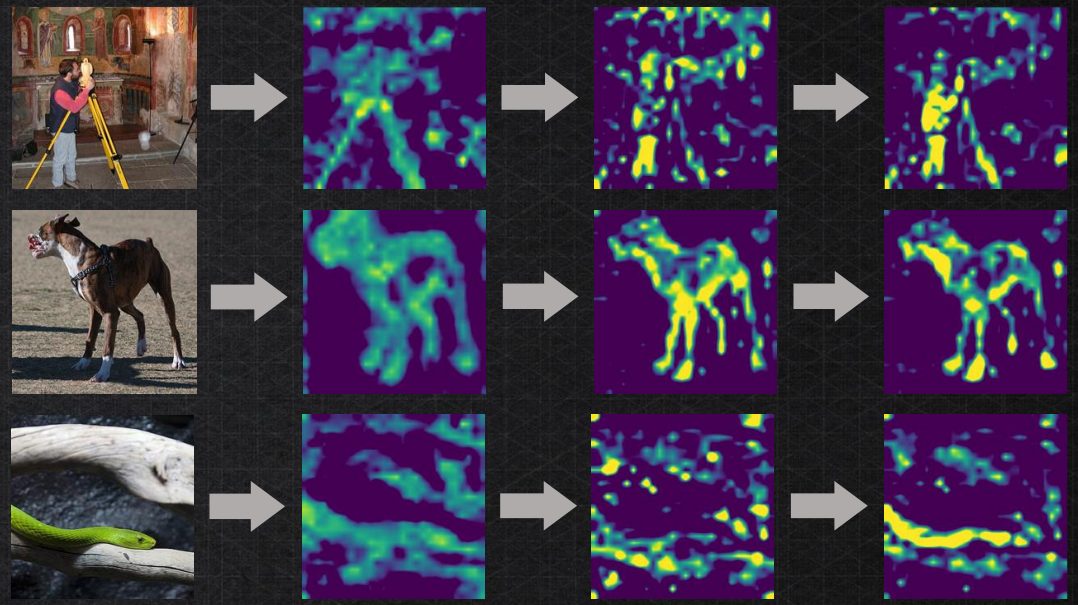
Baifeng Shi

University of California, Berkeley

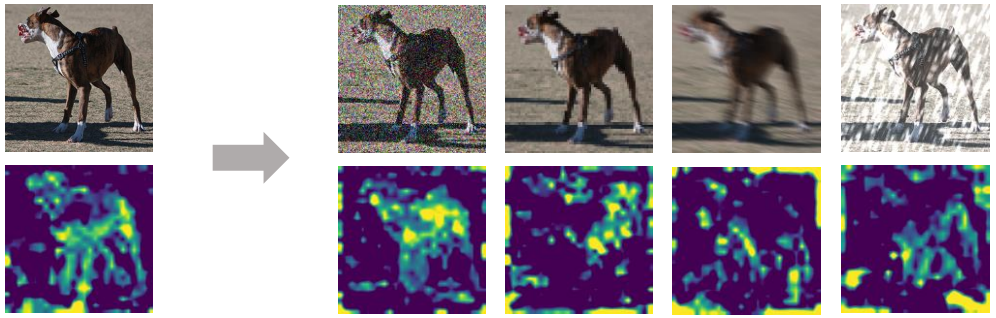
Joint work with Yale Song, Neel Joshi, Trevor Darrell, Xin Wang

Appearing at T4V workshop, CVPR 2022, New Orleans

and ICML 2022, Baltimore



Self-Attention is still Imperfect



- Self-attention based transformers are still not robust enough.
- Here the model failed under image corruption

Attention maps are not matching the intuitive human attention represented by eye fixation maps.

[Visualization of ViT's performance is from the VARS paper.]

VARS: Visual Attention emerges from Recurrent Sparse reconstruction

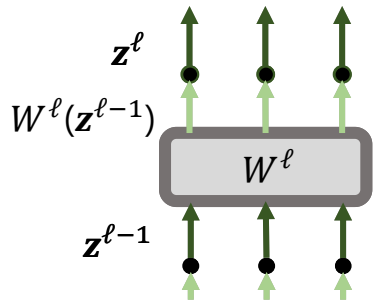
Inspired by two key features in human visual attention: ***recurrency*** and ***sparsity***,

- We propose VARS, Visual Attention from Recurrent Sparse reconstruction, a new attention algorithm
  - which can be plugged into various neural networks to improve model robustness.

- We also show connections between VARS and existing attention algorithms and human attention
  - Self-attention is a special case of VARS.
  - VARS is closer to human attention compared to existing methods.

Adding Recurrency to Feedforward Networks

Feedforward Networks

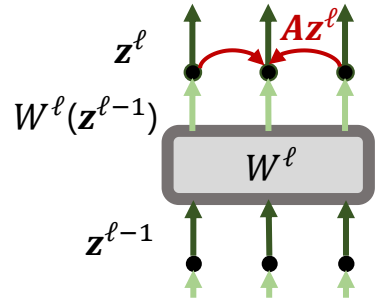


Feed-forward input to neurons

Output of neurons

$$\mathbf{z}^{\ell} = W^{\ell}(\mathbf{z}^{\ell-1})$$

Feedforward Networks w/ recurrency



Feed-forward input to neurons

Recurrent input of neurons

Output of neurons

$$\mathbf{z}^{\ell} = \mathbf{A}^{\ell}\mathbf{z}^{\ell} + W^{\ell}(\mathbf{z}^{\ell-1})$$

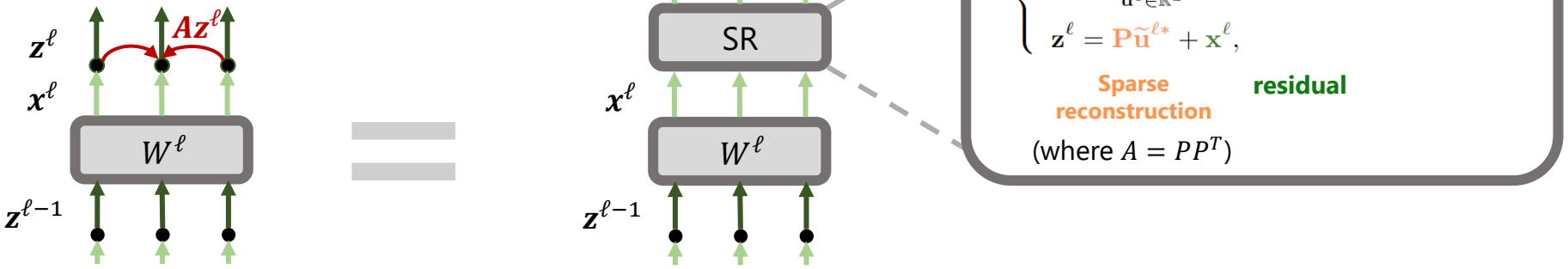
No closed-form solution!

unroll

$$\mathbf{z}_{t+1}^{\ell} = \mathbf{A}^{\ell}\mathbf{z}_{t}^{\ell} + W^{\ell}(\mathbf{z}^{\ell-1})$$

Recurrency Entails Sparse Reconstruction

Recurrently connected network = Feedforward network w/ additional sparse reconstruction layer*.

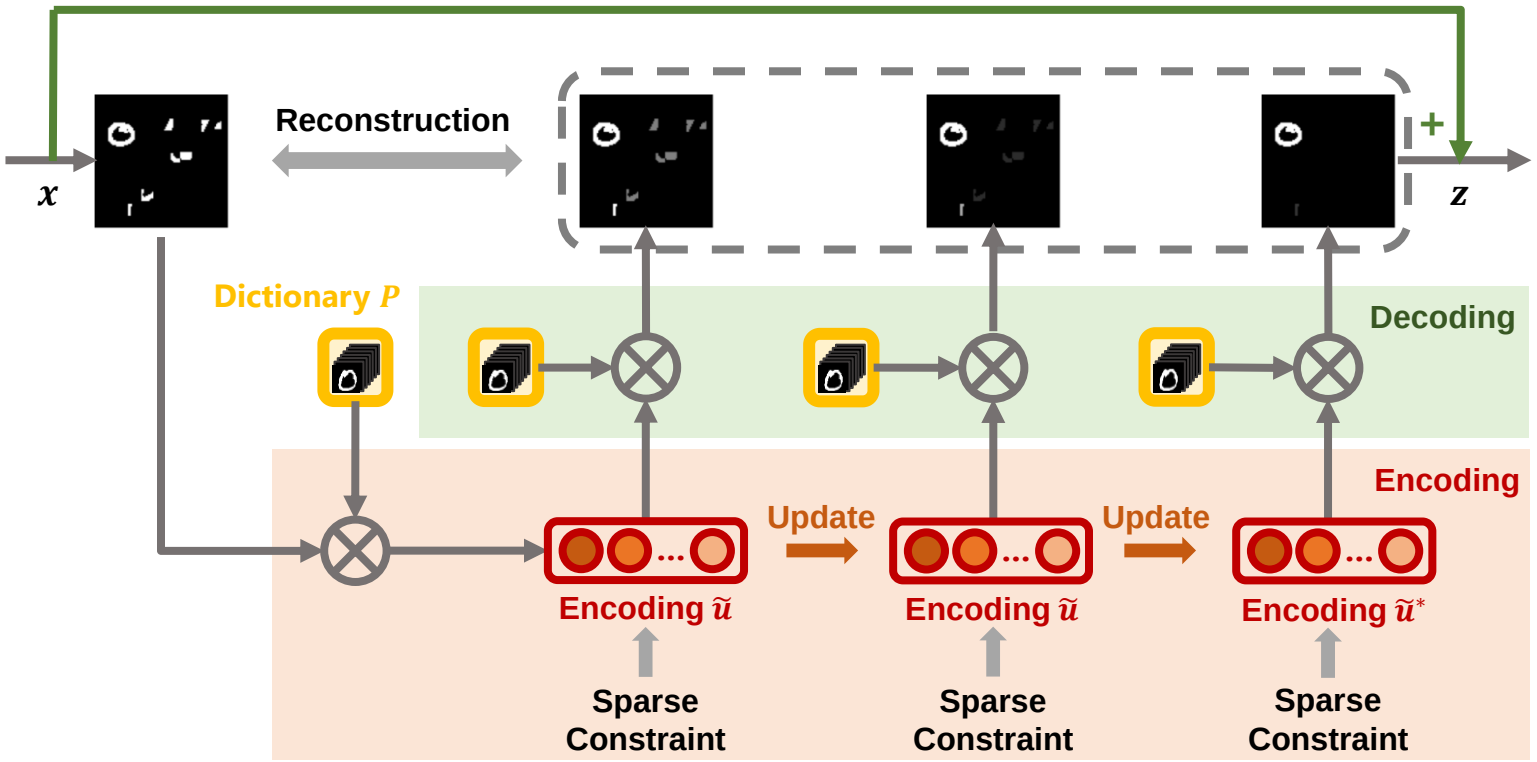


Feedforward network
w/ recurrency

Feedforward network
w/ additional sparse reconstruction layer
(VARS)

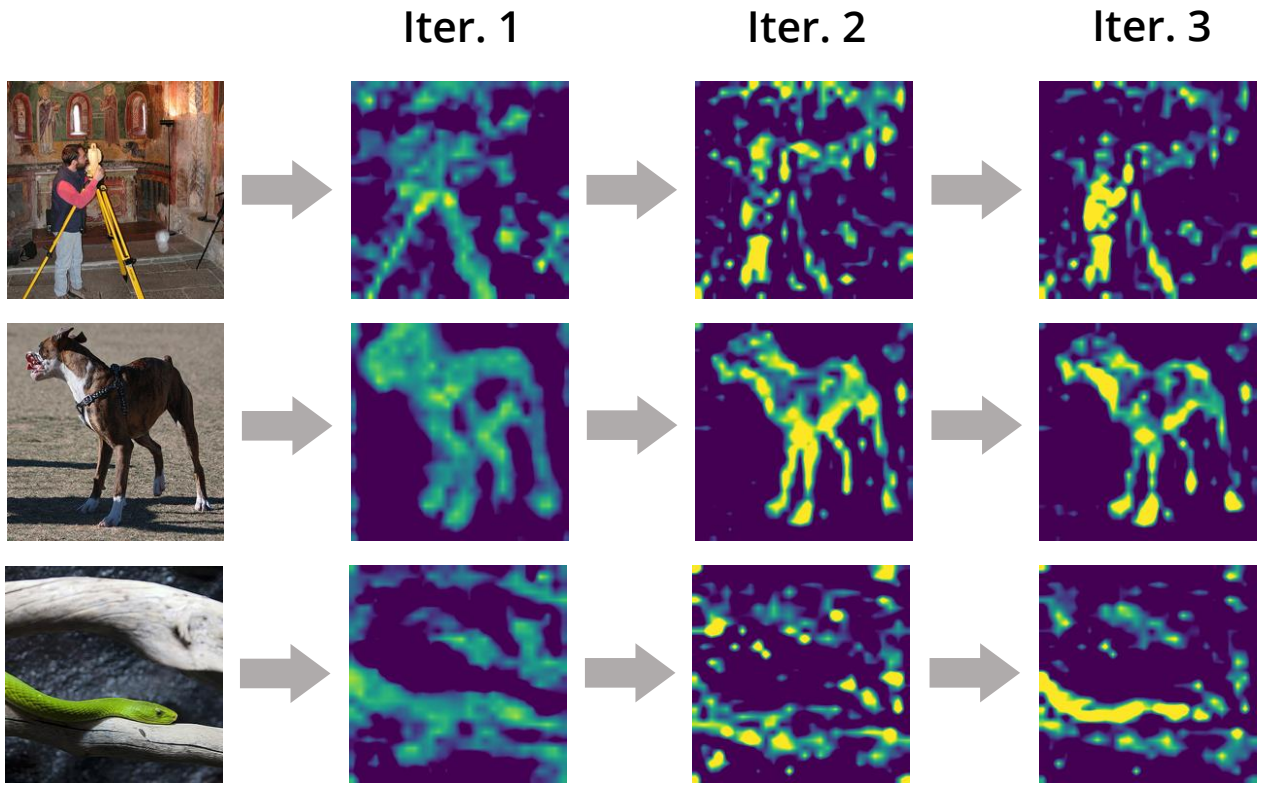
* Under certain conditions.

Attention Emerges from Sparse Reconstruction



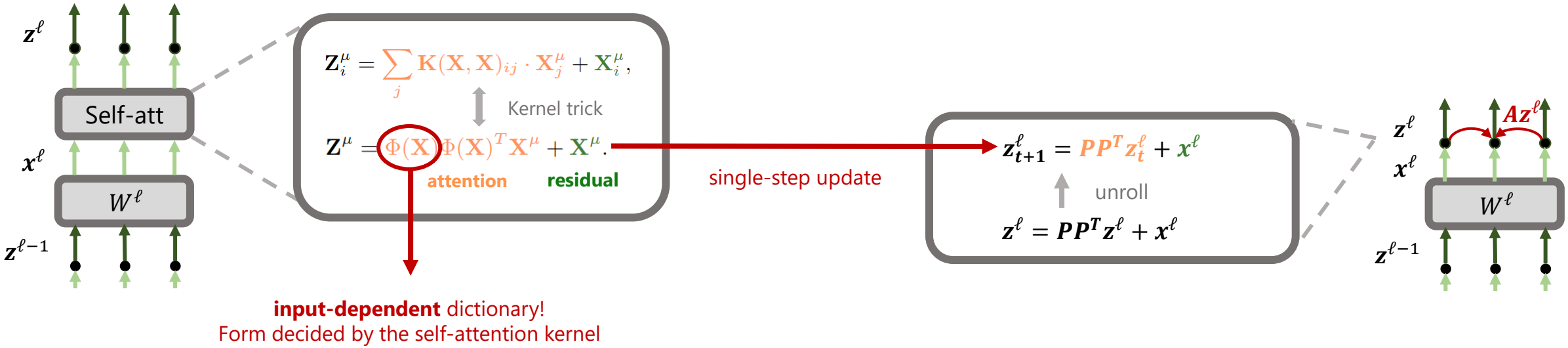
In each VARS block, we iteratively optimize the sparse reconstruction using ISTA (LISTA)

Iterative Improvement of Attention Maps of VARS



The feature map is slowly focused on the salient objects.

Self-Attention as a Special Case of VARS



input-dependent dictionary!
Form decided by the self-attention kernel

Choromanski K, Likhosherstov V, Dohan D, et al. Rethinking attention with performers.

Performance on Robustness Benchmarks

Our models:

- VARS-S: **static** dictionary
- VARS-D: **dynamic** dictionary (similar to self-attention)
- VARS-SD: **static + dynamic**

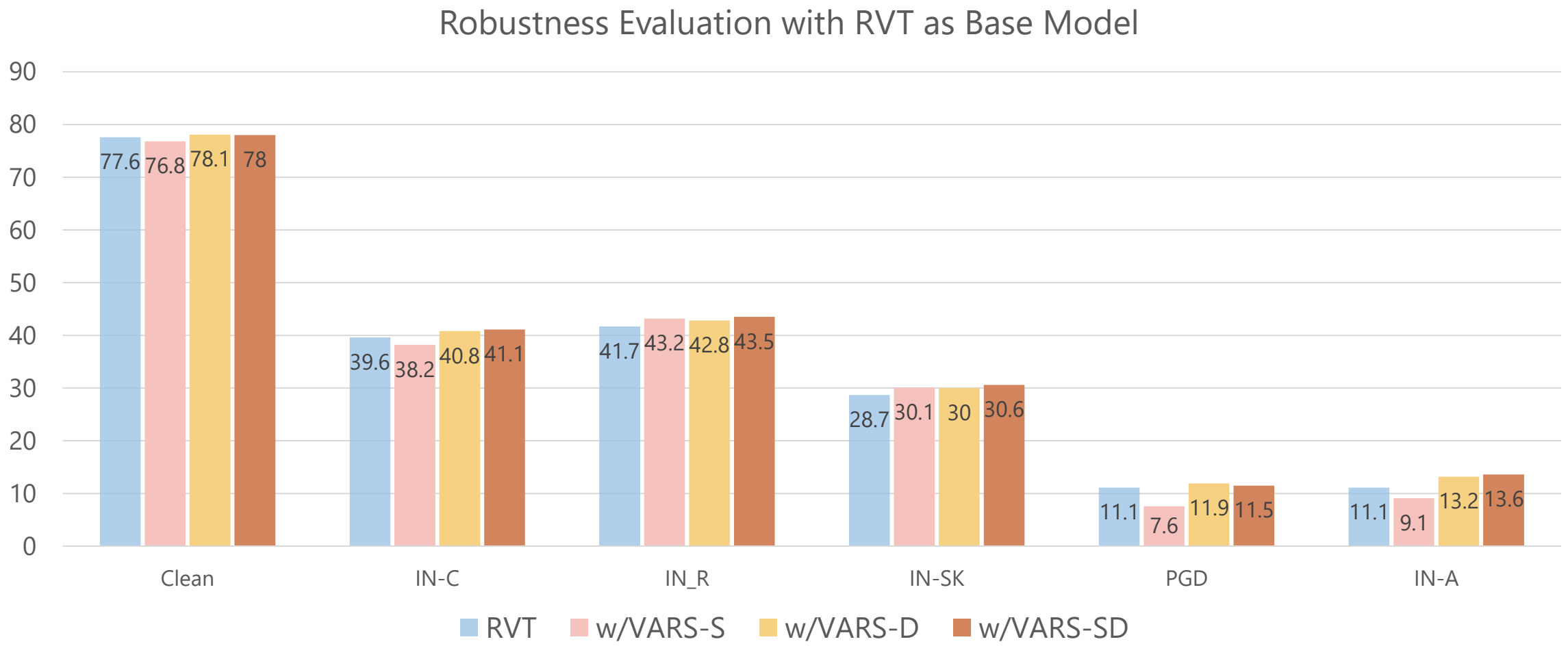
Baseline Architectures:

- DeiT
- RVT (Robust Vision Transformer)
- GFNet (Global Filter Network)

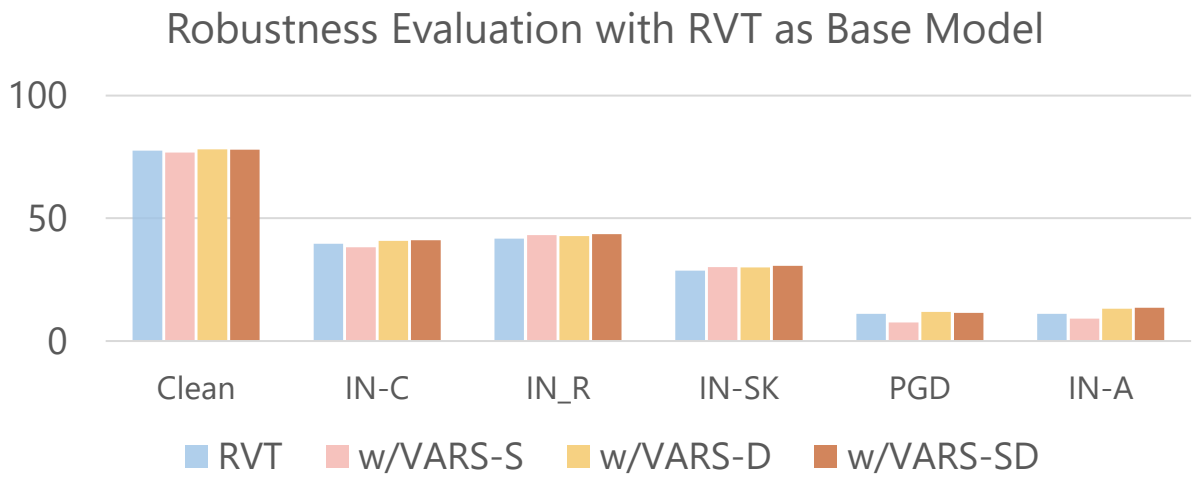
Datasets:

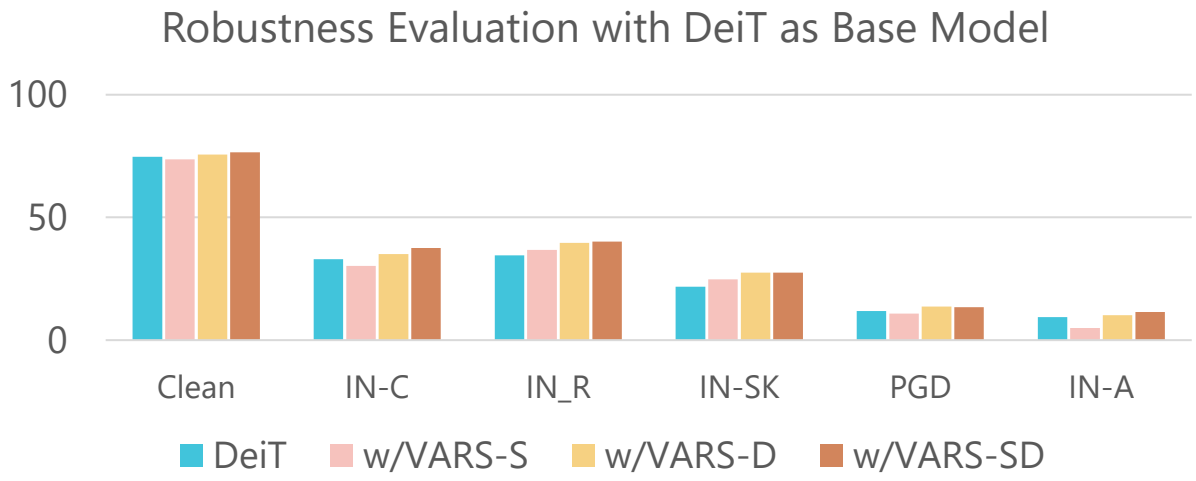
	Dataset Name	Type
Robustness	ImageNet-C	Natural corruption
	ImageNet-R	Out of distribution
	ImageNet-SK	Out of distribution
	PGD	Adversarial attacks
	ImageNet-A	Natural adv. examples
Others	PACS	Domain generalization
	PASCAL VOC	Semantic segmentation
	MIT1003	Human eye fixation

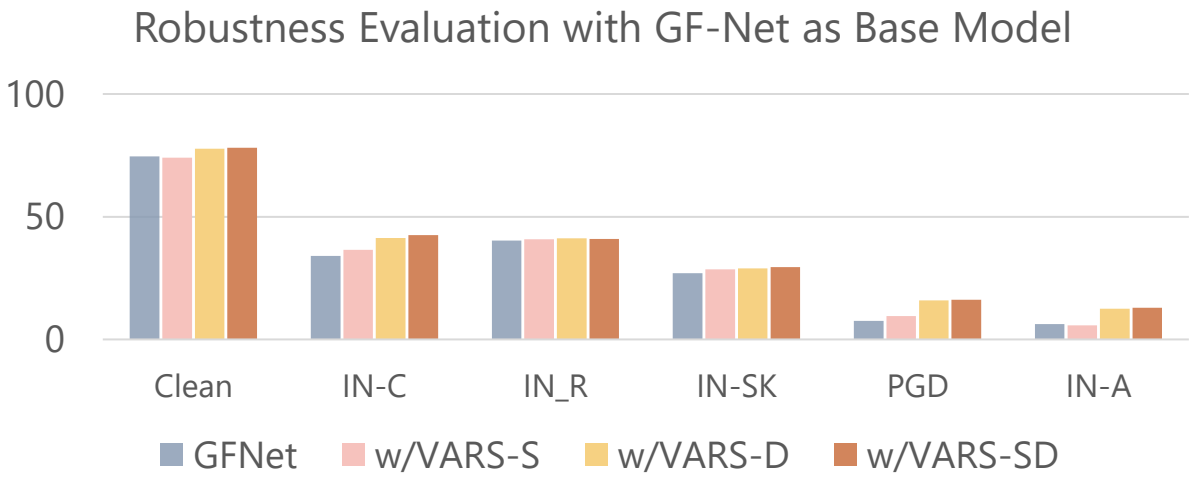
Performance on Robustness Benchmarks



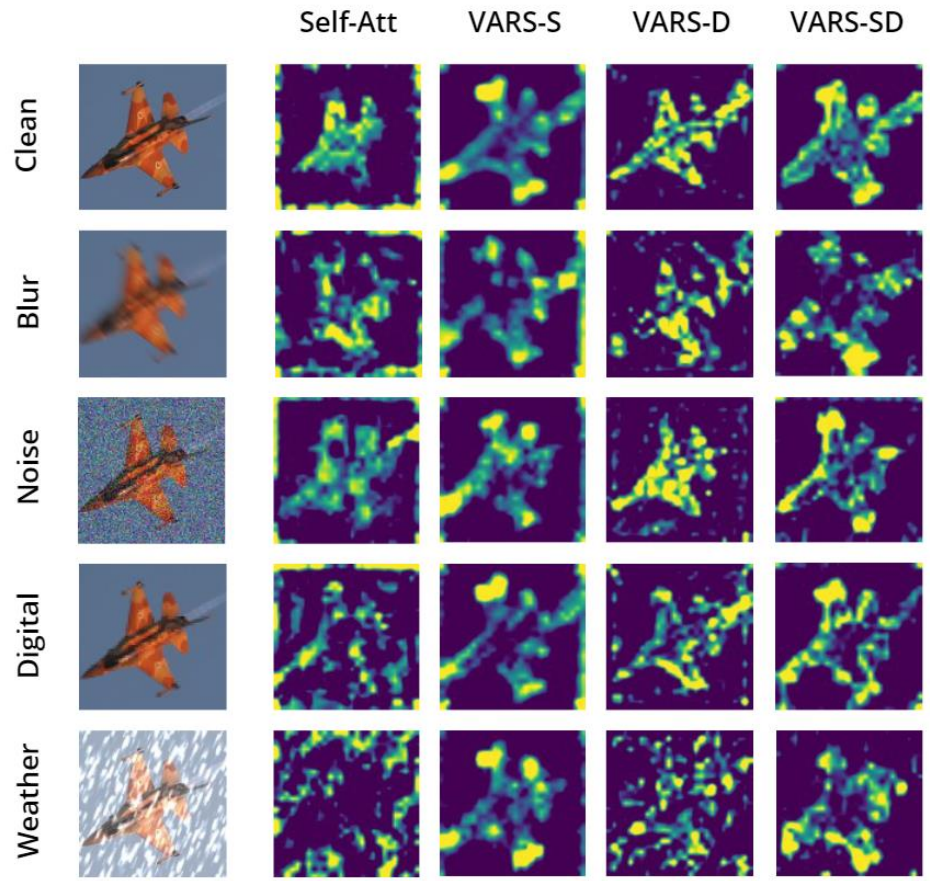
Performance on Robustness Benchmarks





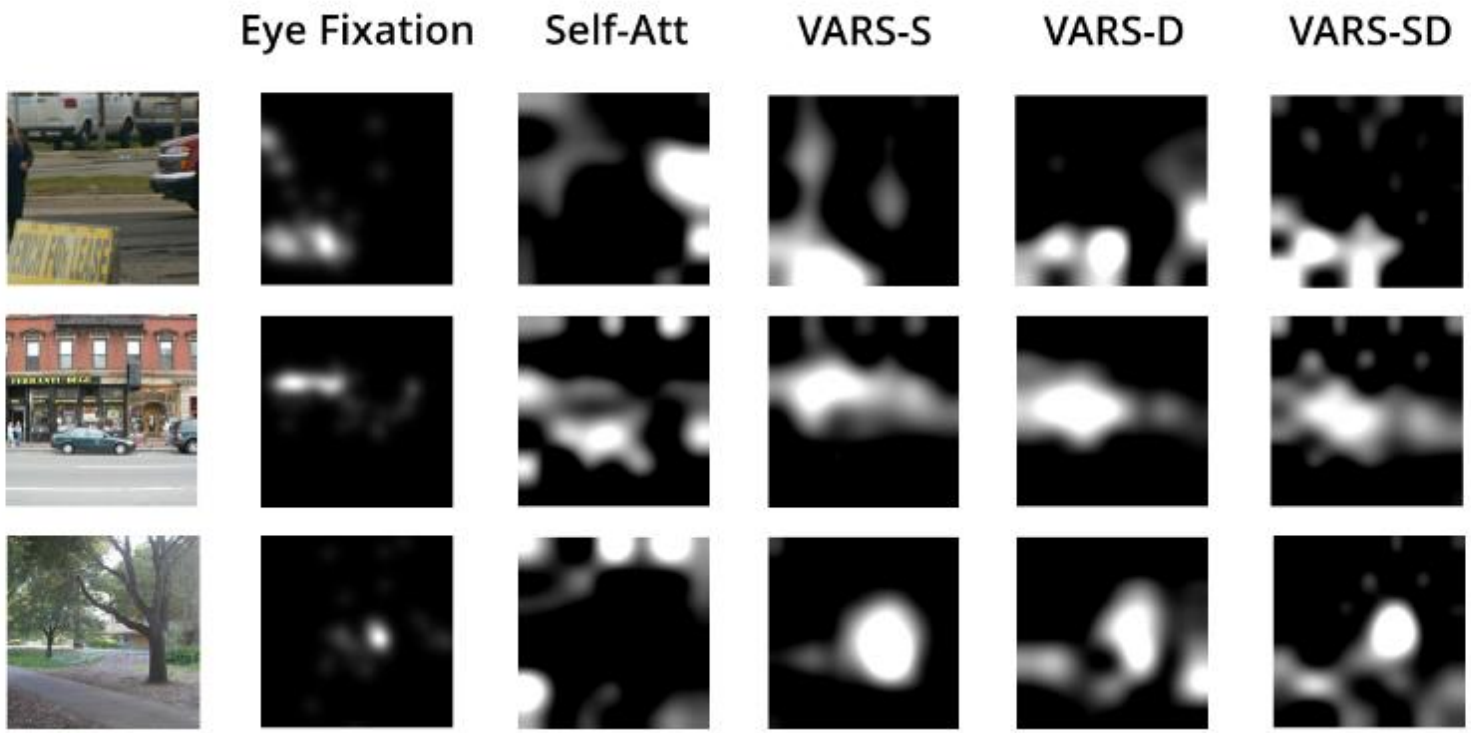


Performance on Robustness Benchmarks



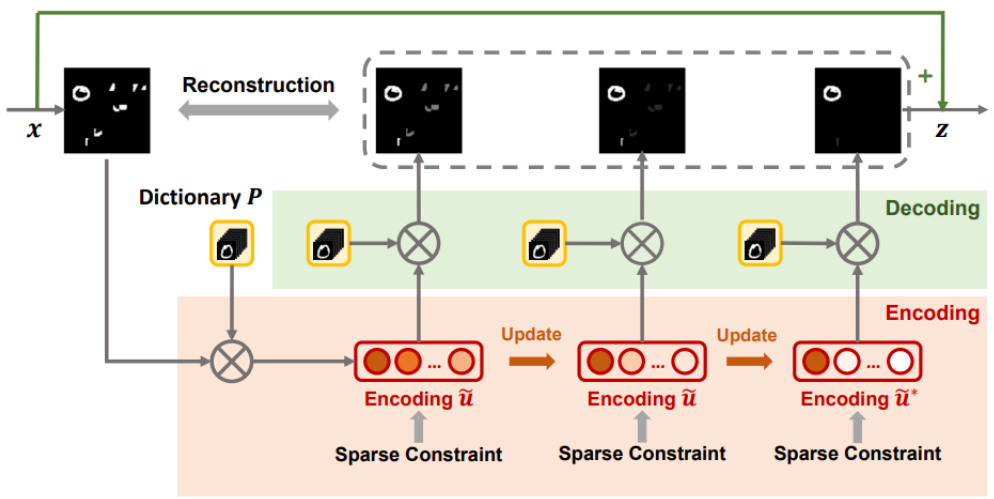
VARS with dynamic dictionaries identify better salient regions that the others.

VARS is More Consistent with Human Eye Fixation



VARS's attention maps are more consistent with human eye fixation than self-attention's.

Summary



- We propose VARS, a new attention design,
  - which can be plugged into various neural networks to improve model robustness.

- We also show that
  - VARS is closer to human attention compared to existing methods.
  - Self-attention is also a special case of VARS.

Thank you!

Baifeng Shi
bfshi.github.io
baifeng_shi@berkeley.edu