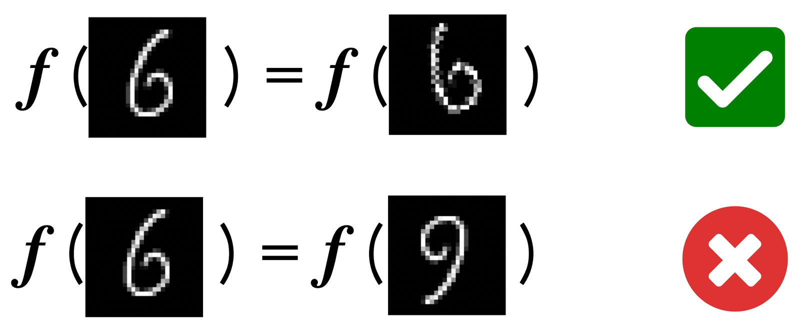
Do Invariances in Deep Neural Networks Align with Human Perception?

Vedant Nanda, Ayan Majumdar, Camila Kolling, John P. Dickerson, Krishna P. Gummadi, Bradley C. Love, Adrian Weller

University of Maryland, College Park · Max Planck Institute for Software Systems · University College London · University of Cambridge

Learning the Right Invariances is Necessary for Robustness

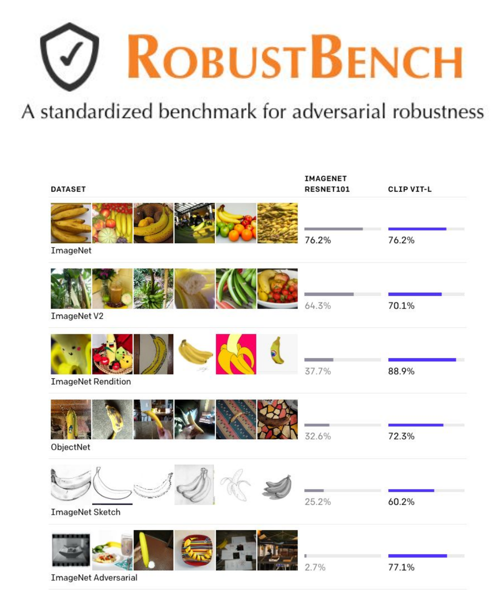


Lack of human-like invariances makes models fail in unexpected ways

In order for models to be robust, they must learn invariances that align with human perception

Ideal Robustness Evaluation

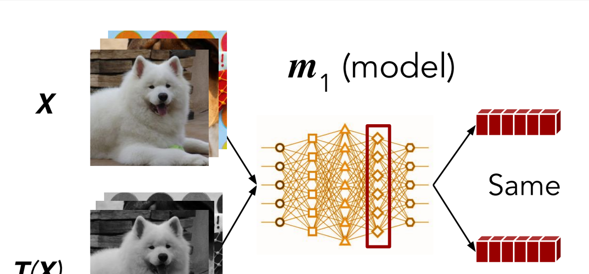
Key Q: How do we measure alignment between model and human invariances?

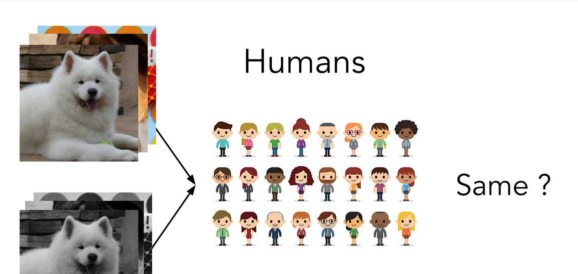


Robustness benchmarks used today are special cases

Eg: adversarial robustness tests invariance to ℓ_p perturbations, ImageNet-C to blur, noise etc.

Our Proposal for Measuring Alignment

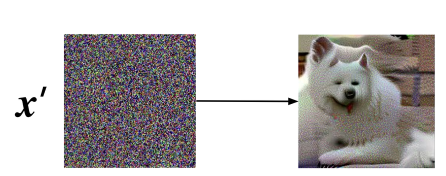


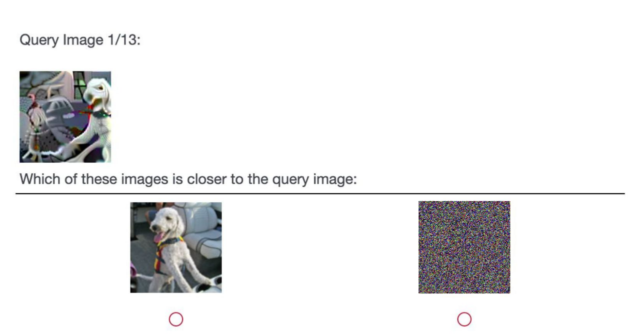


Representation Inversion (Mahendran&Vedaldi, 2015)

Ask human annotators* if x is closer to initial x′ or final x′

$$\mathrm{argmin}_{x'}\ \mathcal{L}(x')$$
$$x' = x' - \alpha\,\nabla_{x'}\mathcal{L}$$





$$\mathcal{L}(x') = \| m_1(x) - m_1(x') \|_2 / \| m_1(x) \|_2 + \lambda^* R(x')$$

*For scalability, we show that we can use perceptual distance LPIPS (Zhang et al., 2018) to automate these judgments

🔥 Our finding:
residual connections +
adversarial data augmentation +
self-supervised training w contrastive loss =
higher alignment 🚀

🔥 We identify challenges and highlight the key role of loss used to generate identically represented inputs in measuring alignment with human perception

🔥 We propose an improved method to measure alignment that does not required labelled data and can scale well

arXiv tinyurl.com/invariances-human

github.com/nvedant07/Human-NN-Alignment

Major Challenges

➢How to choose X? [Results in the paper]

➢No access to humans' representation [Visual similarity]

➢How to pick T? *i.e.* How to define the regularizer R?

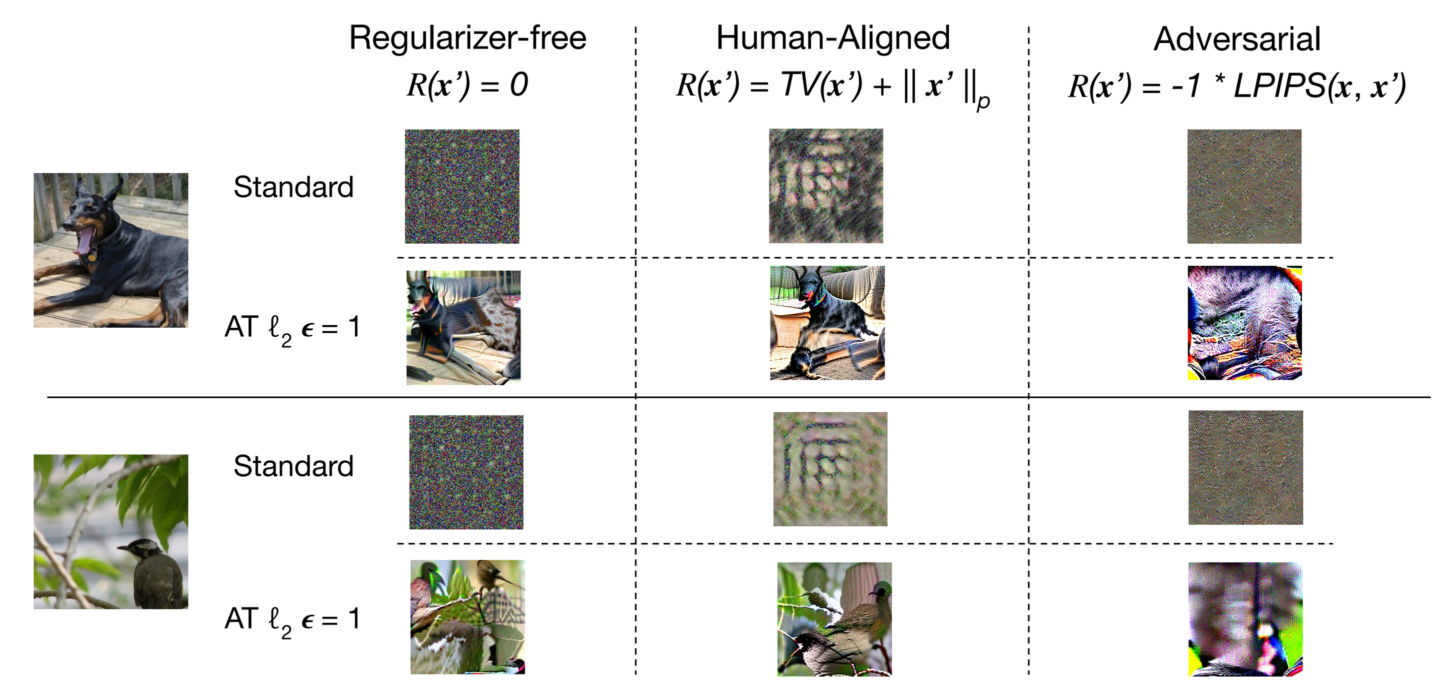
Prior works on measuring alignment have seemingly contradictory takeaways since they do not explicitly engage with the choice of R

(Colah et al.,2017) Certain parts of DNNs encode human-like concepts

(Engstrom et al., 2019) Adversarially trained models induce a "human prior" over learned representations

(Feather et al., 2019) Invariances in later layers diverge from human perception

Importance of Loss Used to Generate Identically Represented Inputs



Training	Model	Alignment: Reg.-Free	Alignment: Human-Aligned	Alignment: Adversarial
CIFAR10				
AT $\ell_2, \epsilon = 1$	ResNet18	$63.25_{\pm 26.23}$	$79.00_{\pm 21.94}$	$0.33_{\pm 0.47}$
	VGG16	$0.25_{\pm 0.43}$	$41.41_{\pm 16.74}$	$1.00_{\pm 1.41}$
	InceptionV3	$23.25_{\pm 25.56}$	$64.75_{\pm 24.17}$	$3.00_{\pm 4.24}$
	DenseNet121	$82.75_{\pm 20.07}$	$86.25_{\pm 14.50}$	$1.33_{\pm 1.89}$
Standard	ResNet18	$0.00_{\pm 0.00}$	$21.09_{\pm 13.51}$	$1.33_{\pm 1.89}$
	VGG16	$0.00_{\pm 0.00}$	$21.88_{\pm 14.82}$	$0.00_{\pm 0.00}$
	InceptionV3	$0.00_{\pm 0.00}$	$21.88_{\pm 17.54}$	$0.33_{\pm 0.47}$
	DenseNet121	$0.00_{\pm 0.00}$	$26.56_{\pm 16.90}$	$0.00_{\pm 0.00}$
ImageNet				
AT $\ell_2, \epsilon = 3$	ResNet18	$42.00_{\pm 38.33}$	$46.75_{\pm 39.37}$	$0.33_{\pm 0.47}$
	ResNet50	$51.00_{\pm 34.89}$	$45.75_{\pm 37.39}$	$14.00_{\pm 3.74}$
	VGG16	$55.50_{\pm 34.14}$	$55.50_{\pm 38.29}$	$11.00_{\pm 3.74}$
Standard	ResNet18	$0.00_{\pm 0.00}$	$17.00_{\pm 28.30}$	$0.00_{\pm 0.00}$
	ResNet50	$0.00_{\pm 0.00}$	$16.25_{\pm 26.42}$	$0.00_{\pm 0.00}$
	VGG16	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$

Olah et al., 2017
Engstrom et al., 2019
Feather et al., 2019

All models are poorly aligned under the pessimistic lens of adversarial regularizer

What Components of DL Pipeline Contribute to Alignment?