Do Invariances in Deep Neural Networks Align with Human Perception?

AAAI 2023

Vedant Nanda
PhD Student
University of Maryland & MPI-SWS

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS


Camila Kolling


Ayan Majumdar


Krishna Gummadi


John Dickerson

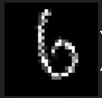

Adrian Weller

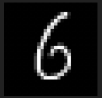
UCL


Bradley Love

Invariances are Crucial for Robust Deep Learning

We need to make sure models learn correct invariances

$f($  $) = f($  $)$

$f($  $) = f($  $)$

Lack of human-like invariances => Models fail in unexpected ways!

Measuring alignment of invariances is a fundamental measure of robustness

Robustness Evaluation Today

Accuracy under adversarial perturbations (Carlini et al., 2019; Madry et al., 2018)

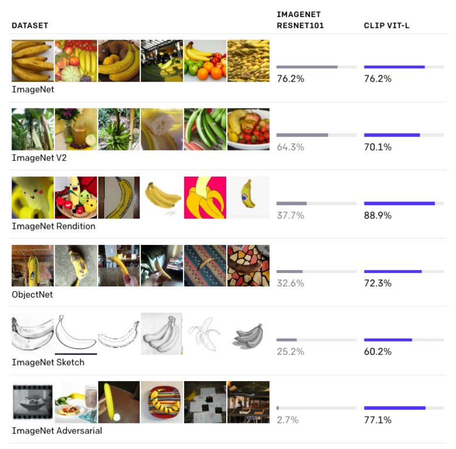
- Evaluate accuracy under worst case perturbation in a given threat model (*eg:* ℓ_p, patch etc)

Accuracy under various distribution shifts

- ImageNetV2 (Recht et al., 2019)
- ImageNet-R (Hendrycks et al., 2021)
- ImageNet-C, ImageNet-P (Hendrycks et al., 2019)
- ObjectNet (Barbu et al., 2019)
- ImageNet-Sketch (Wang et al., 2019)
- ImageNet-A (Hendrycks et al., 2019)

ROBUSTBENCH
A standardized benchmark for adversarial robustness

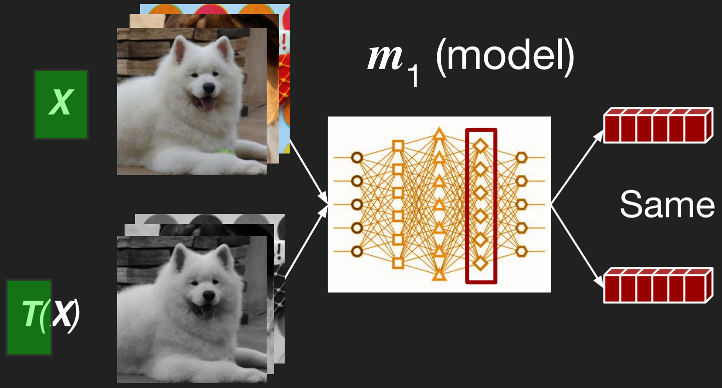
https://robustbench.github.io (Croce et al., 2021)

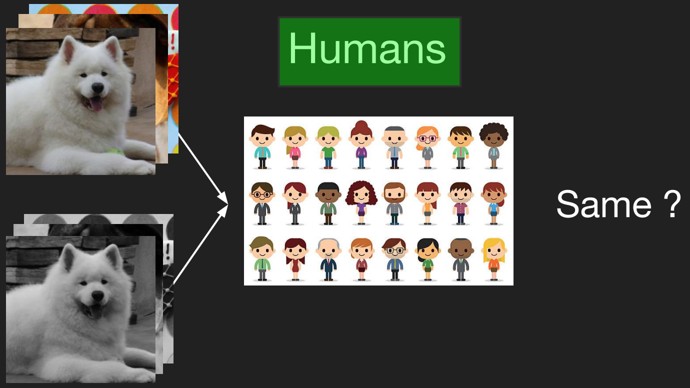


https://openai.com/blog/clip (Radford et al., 2021)

The Other Direction of Robustness Evaluation

Do Invariances in DNNs align with Human Perception?





- How to choose X?
- [Choosing T] Infinitely many T. How to pick appropriate T?
- [Humans] No access to representations in human brain

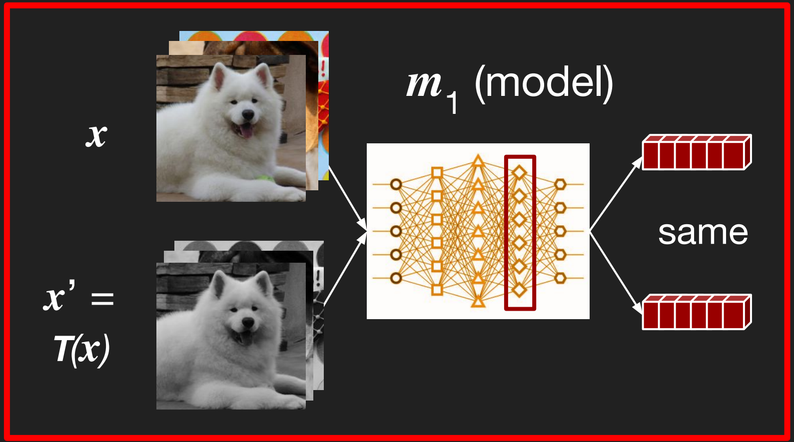
Our Contribution

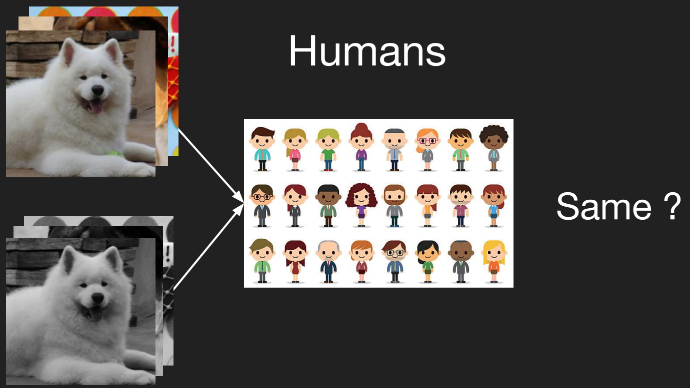
- [Choosing T] Highlight the role of loss function used in finding invariant transforms
    - Reconcile seemingly contradictory takeaways in prior work
- [Humans] Provide an improved way of measuring alignment with human perception
    - Does not require labelled data
    - Scalable
- Analyze how architectures, losses, data augmentations affect alignment

Our Contribution

- [Choosing T] Highlight the role of loss function used in finding invariant transforms
  - Reconcile seemingly contradictory takeaways in prior work
- [Humans] Provide an improved way of measuring alignment with human perception
  - Does not require labelled data
  - Scalable
- Analyze how architectures, losses, data augmentations affect alignment

Robustness Evaluation



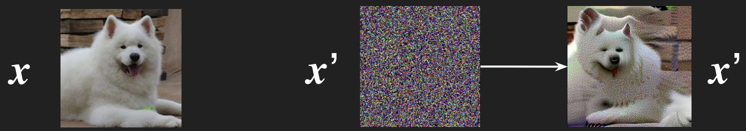


1. Find *Identically Represented Inputs (IRIs)* x, x' such that $m_1(x) \approxeq m_1(x')$

Representation Inversion (Mahendran & Vedaldi, CVPR 2015)

$$\mathit{argmin}_{x'} \boxed{\mathcal{L}(x')}$$

$$x' = x' - \alpha \nabla_{x'} \mathcal{L}$$



s.t. $m_1(x) \approxeq m_1(x')$

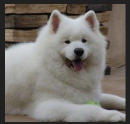
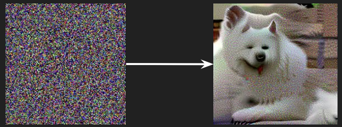
Loss Used to Generate IRIs

1. Find *Identically Represented Inputs (IRIs)* $\boldsymbol{X}$, $\boldsymbol{X'}$ such that $m_1(\boldsymbol{X}) \cong m_1(\boldsymbol{X'})$

Representation Inversion ([Mahendran & Vedaldi, CVPR 2015](#))

$$\text{argmin}_{x'} \boxed{\mathcal{L}(x')}$$

$$x' = x' - \alpha \nabla_{x'} \mathcal{L}$$

x  x'  x'

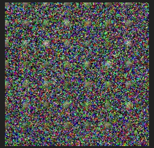
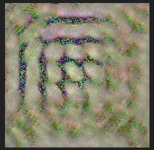
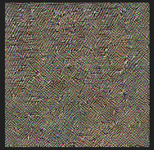
$$\mathcal{L}(x') = \| m_1(x) - m_1(x') \|_2 + \lambda * \boxed{R(x')}$$

Regularizer-free $R(x') = 0$

Human-Aligned $R(x') = TV(x') + \| x' \|_p$ Removes high-frequency components from x'

Adversarial $R(x') = -1 * LPIPS(x, x')$ Makes x and x' perceptually distant ([Zhang et al., 2018](#))

Loss Used to Generate IRIs

		Regularizer-free $R(x') = 0$	Human-Aligned $R(x') = TV(x') + \| x' \|_p$	Adversarial $R(x') = -1 * LPIPS(x, x')$
	Standard			
	AT ℓ_2 $\epsilon = 1$			

Regularizer impacts takeaways about alignment!

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

CIFAR10				
Training	Model	Alignment: Reg.-Free	Alignment: Human-Aligned	Alignment: Adversarial
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				
Training	Model	Alignment: Reg.-Free	Alignment: Human-Aligned	Alignment: Adversarial
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

Parts of DNNs encode human-like concepts

Prior works do not directly engage with the choice of regularizer and hence make incomplete conclusions

Under the pessimistic lens of adversarial regularizer all models are poorly aligned

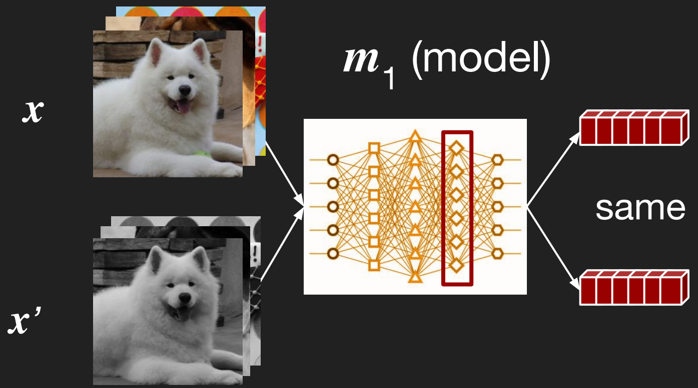
Our Contribution

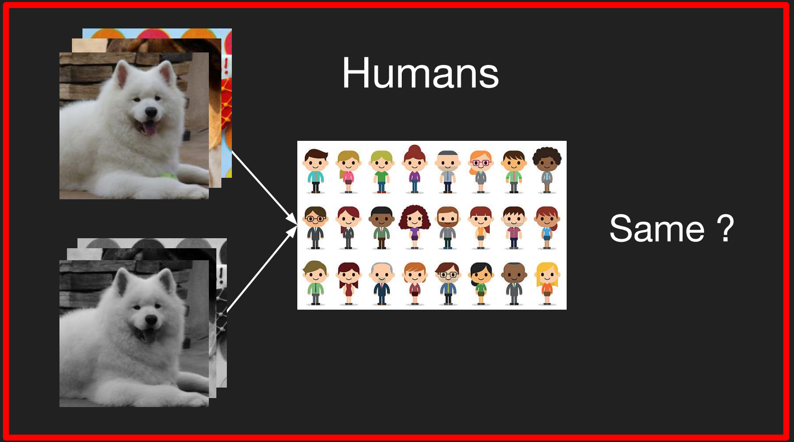
- [Choosing T] Highlight the role of loss function used in finding invariant transforms
  - Reconcile seemingly contradictory takeaways in prior work – choice of regularizer impacts takeaways
- [Humans] Provide an improved way of measuring alignment with human perception
  - Does not require labelled data
  - Scalable
- Analyze how architectures, losses, data augmentations affect alignment

Our Contribution

- [Choosing T] Highlight the role of loss function used in finding invariant transforms
  - Reconcile seemingly contradictory takeaways in prior work – choice of regularizer impacts takeaways
- [Humans] Provide an improved way of measuring alignment with human perception
  - Does not require labelled data
  - Scalable
- Analyze how architectures, losses, data augmentations affect alignment

Robustness Evaluation

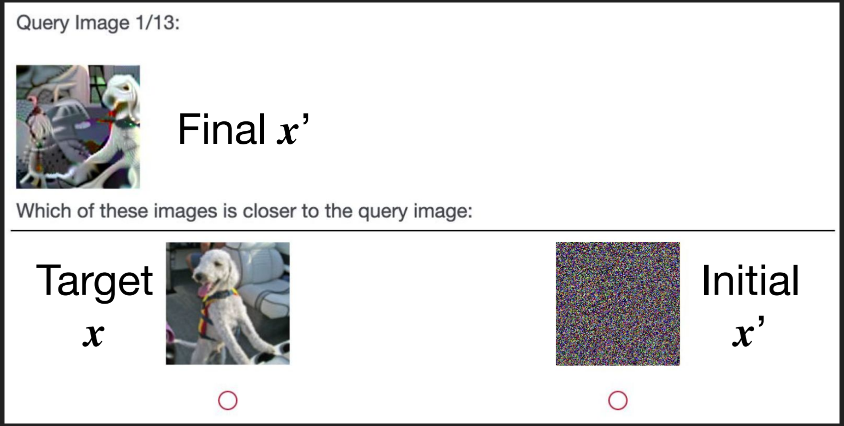




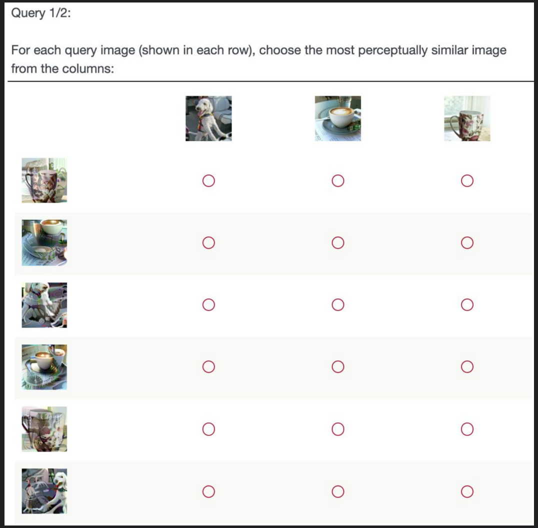
We need to *reliably* and *scalably* check if Humans perceive x and x' similarly

Check if humans perceive these inputs similarly

2AFC



Clustering



Scalability: since these tests are based on comparisons, we can use perceptual distance measures like LPIPS to simulate humans (Zhang et al., 2018)

Evaluation: Reliability

Training	Model	Human 2AFC	Human Clustering
			CIFAR10
Robust	ResNet18	96.00$_{\pm 2.55}$	97.48$_{\pm 1.80}$
	VGG16	38.83$_{\pm 7.59}$	55.387$_{\pm 5.629}$
	InceptionV3	82.00$_{\pm 8.44}$	84.47$_{\pm 6.32}$
	DenseNet121	98.67$_{\pm 0.24}$	97.64$_{\pm 2.08}$
Standard	ResNet18	0.17$_{\pm 0.24}$	38.55$_{\pm 1.19}$
	VGG16	0.17$_{\pm 0.24}$	33.84$_{\pm 2.70}$
	InceptionV3	0.17$_{\pm 0.24}$	38.38$_{\pm 4.06}$
	DenseNet121	9.83$_{\pm 9.97}$	42.42$_{\pm 5.02}$
			IMAGENET
Training	**Model**	**Human 2AFC**	**Human Clustering**
Robust	ResNet18	93.17$_{\pm 5.95}$	96.00$_{\pm 3.59}$
	ResNet50	99.50$_{\pm 0.00}$	99.49$_{\pm 0.71}$
	VGG16	95.50$_{\pm 2.12}$	91.75$_{\pm 5.22}$
Standard	ResNet18	0.00$_{\pm 0.00}$	33.33$_{\pm 0.00}$
	ResNet50	5.33$_{\pm 7.54}$	38.38$_{\pm 2.53}$
	VGG16	0.00$_{\pm 0.00}$	33.96$_{\pm 2.00}$

Both clustering and 2AFC achieve similar ranking among models

Low variance among annotators

=> Humans can determine alignment reliably

Evaluation: Scalability

		CIFAR10			
Training	Model	Human 2AFC	LPIPS 2AFC	Human Clustering	LPIPS Clustering
Robust	ResNet18	$96.00_{\pm 2.55}$	98.0	$97.48_{\pm 1.80}$	91.41
	VGG16	$38.83_{\pm 7.59}$	11.5	$55.387_{\pm 5.629}$	46.97
	InceptionV3	$82.00_{\pm 8.44}$	87.5	$84.47_{\pm 6.32}$	76.77
	DenseNet121	$98.67_{\pm 0.24}$	100.0	$97.64_{\pm 2.08}$	97.98
Standard	ResNet18	$0.17_{\pm 0.24}$	0.0	$38.55_{\pm 1.19}$	31.31
	VGG16	$0.17_{\pm 0.24}$	0.0	$33.84_{\pm 2.70}$	31.31
	InceptionV3	$0.17_{\pm 0.24}$	1.0	$38.38_{\pm 4.06}$	38.38
	DenseNet121	$9.83_{\pm 9.97}$	0.5	$42.42_{\pm 5.02}$	36.87

		ImageNet			
Training	Model	Human 2AFC	LPIPS 2AFC	Human Clustering	LPIPS Clustering
Robust	ResNet18	$93.17_{\pm 5.95}$	82.00	$96.00_{\pm 3.59}$	76.77
	ResNet50	$99.50_{\pm 0.00}$	85.00	$99.49_{\pm 0.71}$	82.83
	VGG16	$95.50_{\pm 2.12}$	82.00	$91.75_{\pm 5.22}$	78.79
Standard	ResNet18	$0.00_{\pm 0.00}$	0.50	$33.33_{\pm 0.00}$	34.85
	ResNet50	$5.33_{\pm 7.54}$	0.50	$38.38_{\pm 2.53}$	35.86
	VGG16	$0.00_{\pm 0.00}$	0.00	$33.96_{\pm 2.00}$	34.34

LPIPS orders models same as humans

=> Can analyze models at scale

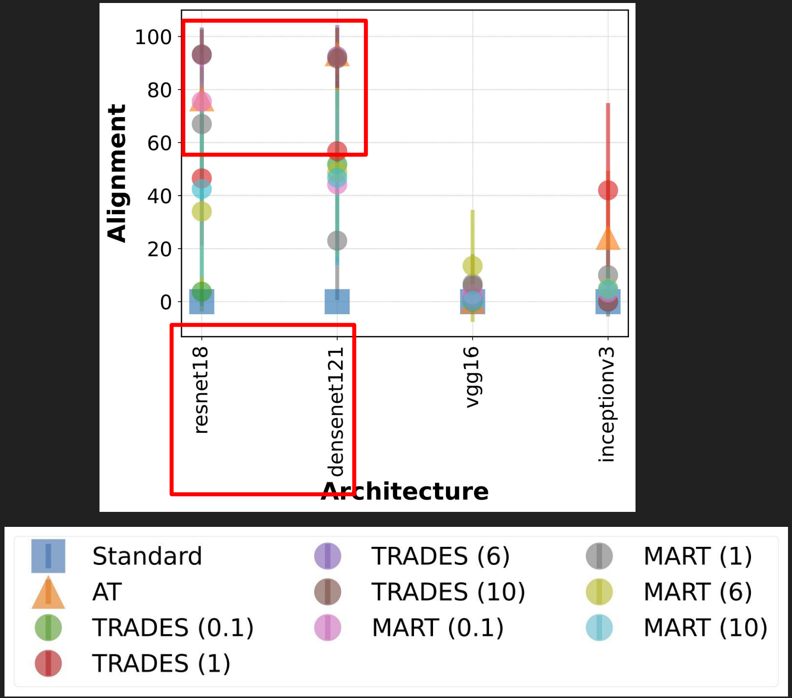
Our Work

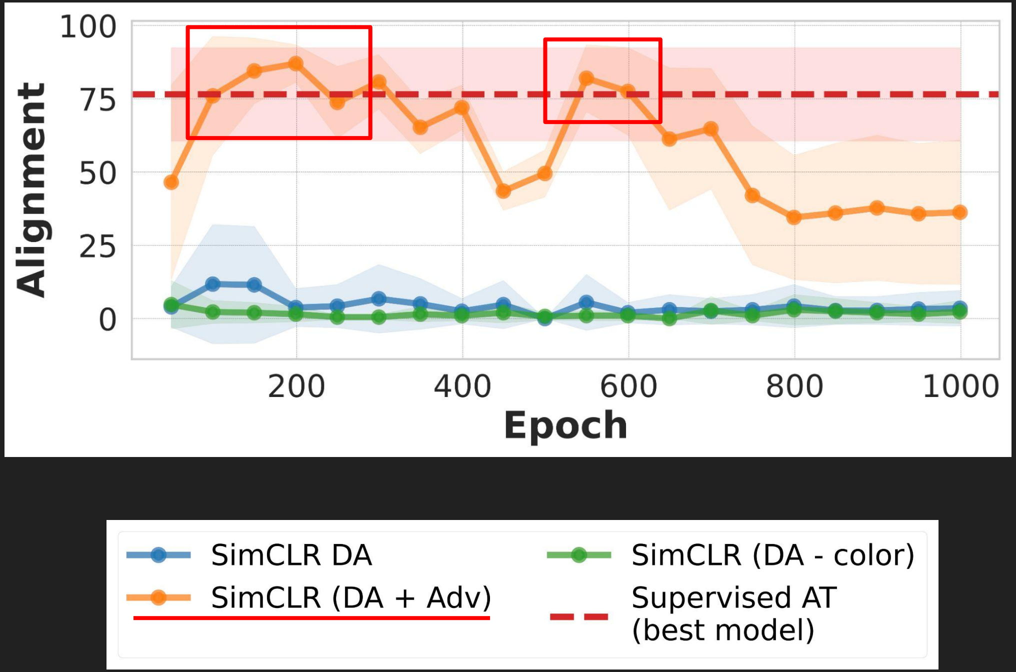
- [Choosing T] Highlight the role of loss function used in finding invariant transforms
  - Reconcile seemingly contradictory takeaways in prior work – choice of regularizer impacts takeaways
- [Humans] Provide an improved way of measuring alignment with human perception
  - Does not require labelled data
  - Scalable
- Analyze how architectures, losses, data augmentations affect alignment

Our Work

- [Choosing T] Highlight the role of loss function used in finding invariant transforms
  - Reconcile seemingly contradictory takeaways in prior work – choice of regularizer impacts takeaways
- [Humans] Provide an improved way of measuring alignment with human perception
  - Does not require labelled data
  - Scalable
- Analyze how architectures, losses, data augmentations affect alignment

Results: Architectures, Losses, Data Augmentations





1. Adversarial data augmentation using ℓ_2 threat model

2. Architectures with residual connections

3. Self-supervised contrastive loss

Summary

- **We highlight challenges and common pitfalls in measuring alignment with human perception**
- **We propose an improved method to measure alignment at scale**
- **Using our method we show how residual connections, adversarial data augmentation and contrastive loss help in increasing alignment**

`tinyurl.com/invariances-human`

`github.com/nvedant07/Human-NN-Alignment`

Thank You!

vnanda@mpi-sws.org

Poster # 111

Prior Works

(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

Contradictory takeaways. What’s going on?

Evaluation

CIFAR10							
Training	Model	Human 2AFC	LPIPS 2AFC	Human Clustering	LPIPS Clustering	Clean Acc.	Robust Acc.
Robust	ResNet18	$96.00_{\pm 2.55}$	98.0	$97.48_{\pm 1.80}$	91.41	80.77	50.92
	VGG16	$38.83_{\pm 7.59}$	11.5	$55.387_{\pm 5.629}$	46.97	79.84	48.36
	InceptionV3	$82.00_{\pm 8.44}$	87.5	$84.47_{\pm 6.32}$	76.77	81.57	51.02
	DenseNet121	$98.67_{\pm 0.24}$	100.0	$97.64_{\pm 2.08}$	97.98	83.22	52.86
Standard	ResNet18	$0.17_{\pm 0.24}$	0.0	$38.55_{\pm 1.19}$	31.31	94.94	0.00
	VGG16	$0.17_{\pm 0.24}$	0.0	$33.84_{\pm 2.70}$	31.31	93.63	0.00
	InceptionV3	$0.17_{\pm 0.24}$	1.0	$38.38_{\pm 4.06}$	38.38	94.59	0.00
	DenseNet121	$9.83_{\pm 9.97}$	0.5	$42.42_{\pm 5.02}$	36.87	95.30	0.00
ImageNet							
Training	Model	Human 2AFC	LPIPS 2AFC	Human Clustering	LPIPS Clustering	Clean Acc.	Robust Acc.
Robust	ResNet18	$93.17_{\pm 5.95}$	82.00	$96.00_{\pm 3.59}$	76.77	53.12	31.02
	ResNet50	$99.50_{\pm 0.00}$	85.00	$99.49_{\pm 0.71}$	82.83	62.83	38.84
	VGG16	$95.50_{\pm 2.12}$	82.00	$91.75_{\pm 5.22}$	78.79	56.79	34.46
Standard	ResNet18	$0.00_{\pm 0.00}$	0.50	$33.33_{\pm 0.00}$	34.85	69.76	0.01
	ResNet50	$5.33_{\pm 7.54}$	0.50	$38.38_{\pm 2.53}$	35.86	76.13	0.00
	VGG16	$0.00_{\pm 0.00}$	0.00	$33.96_{\pm 2.00}$	34.34	73.36	0.16

Our Work

- Reconcile difference in takeaways of prior work [Choosing T]
    - Highlight challenges and common pitfalls in measuring alignment – choice of regularizer impacts takeaways
- Provide an improved way of measuring alignment with human perception [Humans]
    - Does not require labelled data
    - Scalable
- Analyze how architectures, losses, data augmentations affect alignment
    - Architectures with residual connections,
    - Adversarial data augmentation using ℓ_2 threat model
    - Self-supervised contrastive loss