Do Invariances in Deep Neural Networks Align with Human Perception?

Vedant Nanda[1,2], Ayan Majumdar[2], Camila Kolling[2], John P. Dickerson[1], Krishna P. Gummadi[2], Bradley C. Love[3,4], Adrian Weller[3,5]

[1]University of Maryland, College Park, USA
[2]Max Planck Institute for Software Systems (MPI-SWS), Germany
[3]The Alan Turing Institute, London, England
[4]University College London, London, England
[5]University of Cambridge, Cambridge, England
{vnanda, ayanm, ckolling, gummadi}@mpi-sws.org, john@cs.umd.edu, b.love@ucl.ac.uk, adrian.weller@eng.cam.ac.uk

Abstract

An evaluation criterion for safe and trustworthy deep learning is how well the invariances captured by representations of deep neural networks (DNNs) are shared with humans. We identify challenges in measuring these invariances. Prior works used gradient-based methods to generate *identically represented inputs* (IRIs), *i.e.*, inputs which have identical representations (on a given layer) of a neural network, and thus capture invariances of a given network. One necessary criterion for a network's invariances to align with human perception is for its IRIs look "similar" to humans. Prior works, however, have mixed takeaways; some argue that later layers of DNNs do not learn human-like invariances yet others seem to indicate otherwise. We argue that the loss function used to generate IRIs can heavily affect takeaways about invariances of the network and is the primary reason for these conflicting findings. We propose an *adversarial* regularizer on the IRI-generation loss that finds IRIs that make any model appear to have very little shared invariance with humans. Based on this evidence, we argue that there is scope for improving models to have human-like invariances, and further, to have meaningful comparisons between models one should use IRIs generated using the *regularizer-free* loss. We then conduct an in-depth investigation of how different components (*e.g.* architectures, training losses, data augmentations) of the deep learning pipeline contribute to learning models that have good alignment with humans. We find that architectures with residual connections trained using a (self-supervised) contrastive loss with ℓ_p ball adversarial data augmentation tend to learn invariances that are most aligned with humans. Code: github.com/nvedant07/Human-NN-Alignment. **We strongly recommend reading the arxiv version of this paper:** https://arxiv.org/abs/2111.14726.

1 Introduction

The ability to train deep neural networks (DNNs) which learn useful features and representations is key for their widespread use (LeCun, Bengio, and Hinton 2015). In domains where DNNs are used for tasks that previously required human intelligence (*e.g.* image classification) and where safety and trustworthiness are important considerations, it is helpful to assess the alignment of the learned representations with human perception. Such assessments can help in understanding and diagnosing issues such as lack of robustness to distribution shifts (Recht et al. 2019), adversarial attacks (Goodfellow, Shlens, and Szegedy 2015) or using undesirable features for a downstream task (Buolamwini and Gebru 2018).

One test of human-machine alignment is whether different images that map to identical internal network representation are also judged as identical by humans. To study alignment with human perception, prior works have used the approach of *representation inversion* (Mahendran and Vedaldi 2014). The key idea is the following: given an input to a neural network, the approach first finds *identically represented inputs* (IRIs), *i.e.* inputs which have similar representations on some given layer(s) of the neural network. In the second step, the inputs that are perceived similarly by the neural network are checked by humans for visual similarity. Thus, the approach relies on estimating whether a transformation of the inputs which is representation invariant to a neural network is also an invariant transformation to the human eye, *i.e.* it checks whether models and humans have shared or aligned invariances.

Prior works use gradient-based methods to generate IRIs for a given target input starting with a random seed input. These works revealed exciting insights: (a) Feather et al. studied representational invariance for different layers of DNNs trained over ImageNet data (using the standard cross-entropy loss). They showed that while later layer representations of DNNs do not share any invariances with human perception, the earlier layers are somewhat better aligned with human perception (Feather et al. 2019). (b) Engstrom et al. found that, unlike standard DNNs, adversarially robust DNNs, *i.e.*, DNNs trained using adversarial training (Madry et al. 2019), learn representations that are well aligned with human perception, even in later layers (Engstrom et al. 2019b). This was also confirmed by other works (Kaur, Cohen, and Lipton 2019; Santurkar et al. 2019). However, some of these findings are contradicted when differently regularized methods are used for generating IRIs, which show that even later layers of DNNs learn human aligned representations (Mahendran and Vedaldi 2014; Olah, Mordvintsev, and Schubert 2017).

We seek to make sense of these confusing earlier results, and thereby to better understand alignment. We show that when we evaluate alignment of DNNs' invariances and human perception using IRIsgenerated using different loss functions, we can arrive at very different conclusions. For ex-

ample, Fig 1 shows how visual similarity of IRIs can vary massively across different categories of losses.

We group existing IRI generation processes into two broad categories: *regularizer-free*, where the goal is to find an IRI without any additional constraints; and *human-leaning*, where the goal is to find an IRI that is also visually human-comprehensible. Additionally, we propose and explore a new (third) broad category, *adversarial*, where the goal is to find an IRI that is visually (from a human perception perspective) far apart from the target input.

We find that compared to the regularizer-free IRI generation approach, the human-leaning IRI generation approach applies strong constraints on the kind of IRIs generated and thus limits the ability to freely explore the large space of possible IRIs. On the other hand, our proposed adversarial approach shows that in the worst case, all models have close to zero alignment, suggesting that there is scope for improvement in designing models that have human-like invariances (as shown in Fig 1 and Table 2). Based on this evidence, we argue that in order to have meaningful comparisons between models, one should measure alignment using the regularizer-free loss for IRI generation.

Many prior works do not formally define a measure that can quantify alignment with human perception beyond relying on visual inspection of the images by the authors (*e.g.* (Olah et al. 2020)). We show how alignment can be quantified reliably by designing simple visual perception tests that can be crowdsourced, *i.e.* used in human surveys. We also show how one can leverage widely used measures of perceptual distance (Zhang et al. 2018) to automate our human surveys, which allows us to obtain insights at a scale not possible in previous works.

Next, inspired by the prior works that suggest that changes in the model training pipeline (as in training adversarially robust DNNs (Engstrom et al. 2019b; Kaur, Cohen, and Lipton 2019)) can lead to human-like invariant representations, we conduct an in-depth investigation to understand which parts of the deep learning pipeline are critical in helping DNNs better learn human-like invariances. We find that certain choices in the deep learning pipeline can significantly help learn representation that have human-like invariances. For example, we show that residual architectures (*e.g.*, ResNets (He et al. 2016)), when trained with a self-supervised contrastive loss (*e.g.*, SimCLR (Chen et al. 2020a)), using ℓ_2 ball adversarial data augmentations (*e.g.*, as in RoCL (Kim, Tack, and Hwang 2020)); the learned representations – while typically having lower accuracies than their fully supervised counterparts – have higher alignment of invariances with human perception. We highlight the following contributions:

- We show how different losses used for generating IRIs lead to different conclusions about a model's shared invariances with human perception, thus leading to seemingly contradictory findings in prior works.
- We propose an adversarial IRI generation loss, using which we show empirically that we can almost always discover invariances of DNNs that do not align with human perception, thus suggesting that there is scope to design better mechanisms to learn representations that are more aligned with human perception.
- We conduct an in-depth study of how loss functions, architectures, data augmentations and training paradigms lead to learning human-like shared invariances.

2 Measuring Shared Invariance with Human Perception

Measuring the extent to which invariances learned by DNNs are shared by humans is a two step process. We first generate IRIs, *i.e.*, inputs that are mapped to identical representations by the DNN. IRIs give us an estimate about the invariances of the DNN. Then, we assess if these inputs are also considered identical by humans. More concretely, if invariances of a given DNN (g_{model}) are shared by humans (g_{human}) on a set of n d-dimensional samples $X \in \mathbb{R}^{n \times d}$, then:

$$g_{\text{human}}(X^i) \approx g_{\text{human}}(X^j) \, \forall \, (X^i, X^j) \in \mathcal{S} \times \mathcal{S} \; ;$$
$$\mathcal{S} = \{X\} \cup \{X^i \mid g_{model}(X^i) \approx g_{model}(X)\}.$$

$\mathcal{S}$ denotes the IRIs for g_{model}. There are three major challenges here:

- Access to representations in the brain, *i.e.*, g_{human} is not available.
- Due to the highly non-linear nature of DNNs, $\mathcal{S}$ can be very hard to obtain.
- The fine-grained input space implies very many inputs n, making the choice of X hard.

We address each of these below. We also show how prior works that do not directly engage with these points can miss important issues in their conclusions about shared invariances of DNNs and humans.

2.1 Approximating g_{human}

Assuming we have a set of images with identical representations ($\mathcal{S}$; how we obtain this is discussed in Section 2.2), we must check if humans also perceive these images to be identical. The extent to which humans think this set of images is identical defines how aligned the invariances learned by the DNN are with human perception. In prior works this has been done by either eyeballing IRIs (Engstrom et al. 2019b) or by asking annotators to assign class labels to IRIs (Feather et al. 2019); both approaches do not scale well. Additionally, assigning class labels to IRIs limits X to being samples from a data distribution containing human-recognizable images (*i.e.*, X cannot be sampled from any arbitrary distribution) with only a few annotations (*e.g.*, asking annotators to assign one class label out of 1000 ImageNet classes is not feasible). To address the issues of scalability and class labels, we propose the following as a measure of alignment between DNN and human invariances:

$$\text{Alignment} = \frac{|\mathcal{A}|}{\sum_{x_t \in X} |\mathcal{S}_{x_t}|}, \text{where} \tag{1}$$

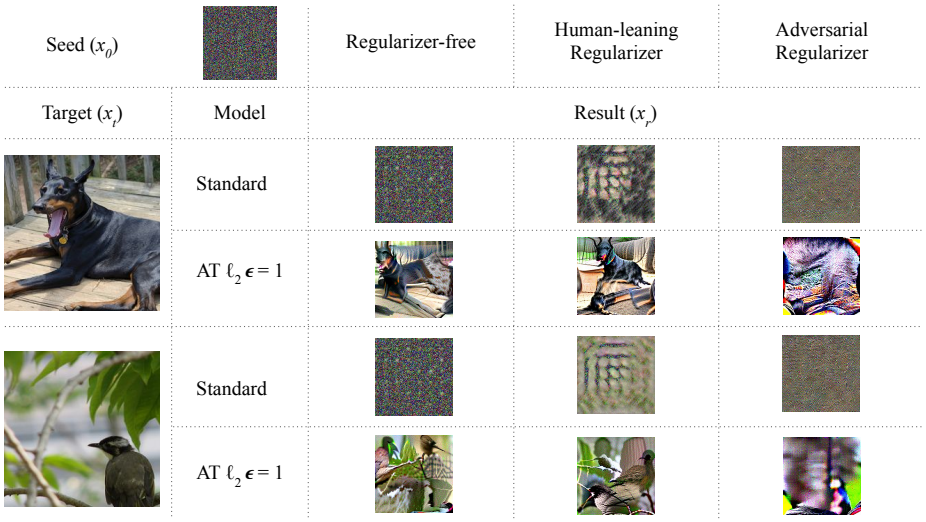
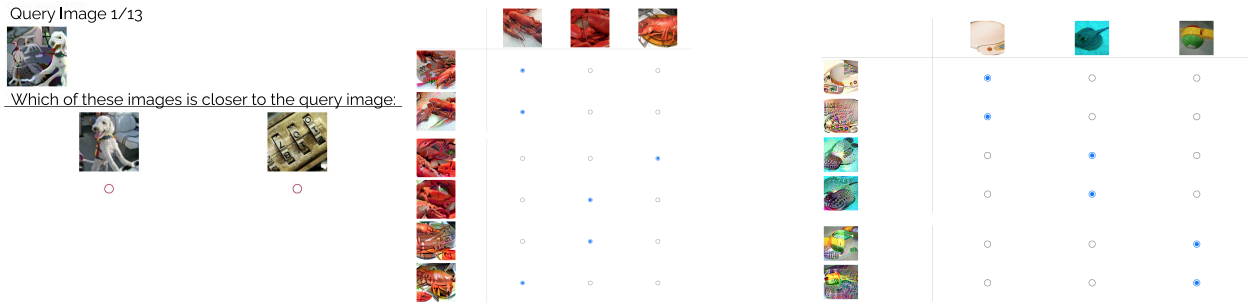


Figure 1: [Representation Inversion for different kinds of $\mathcal{R}$; For ImageNet trained ResNet50] For the standard ResNet50, with regularizer-free and adversarial inversion, x_r looks perceptually much closer to x_0 than x_t, even though from the model's point of view, x_r and x_t are the same. However, with the human-leaning regularizer, we see that x_r contains some information like color patterns of x_t. For adversarially robust ResNet50 (Salman et al. 2020) even though regularizer-free and human-leaning inversions look perceptually similar to x_t, for the adversarial regularizer even these models produce x_r that looks nothing like x_t. Images are generated by starting from x_0 and solving Eq 2 with different kinds of regularizers.



(a) 2AFC (b) Hard ImageNet Clustering (c) Random ImageNet Clustering

Figure 2: [Survey Prompts for AMT workers] In the 2AFC (left) setting we ask the annotator to choose which of the two images (x_t or x_0) is perceptually closer to the query image (x_r). In the clustering setting (center and right) we show 3 images from the dataset (target images, x_t) in the columns and for each of these, we generate $x_{r_1} \in \mathcal{S}_{x_t}$ and $x_{r_2} \in \mathcal{S}_{x_t}$. Each of these is shown across the rows. The task here is to match each image on the row with the corresponding target image on the column.

$$\mathcal{A} = \{x_{r_i} |\ ||g_{\text{human}}(x_t) - g_{\text{human}}(x_{r_i})|| < ||g_{\text{human}}(x_0) - g_{\text{human}}(x_{r_i})||\ \ \forall x_t \in X, x_{r_i} \in S_{x_t}\},$$
$$S_{x_t} = \{x_{r_i} | g_{\text{model}}(x_{r_i}) \approx g_{\text{model}}(x_t)\ \ \forall x_t \in X\},$$

where x_0 is the starting point for Eq 2 sampled from $\mathcal{N}(0, 1)$. In Section 2.4 we see how alignment is robust to the choice of x_0. By directly looking for perceptual similarity of IRIs (captured by $\mathcal{A}$), we get past the issue of assigning class labels to IRIs. The comparison used to generate $\mathcal{A}$ is referred to as the 2 alternative forced choice test (2AFC) which is commonly used to assess sensitivity of humans to stimuli (Fechner, Howes, and Boring 1966). In order to compute $\mathcal{A}$, we estimate perceptual distance $d(x_i, x_j) = ||g_{\text{human}}(x_i) - g_{\text{human}}(x_j)||$ between two inputs. Ideally, we would like to measure $d(x_i, x_j)$ by directly asking for human annotations, however, this approach is expensive and does not scale when we wish to evaluate many models. To address scalability, we use LPIPS (Zhang et al. 2018) which is a commonly used measure for perceptual distance and thus can be used to approximate $d(x_i, x_j)$ [1]. While LPIPS is by no means a perfect approximation, it allows us to gain insights at a scale not possible in prior works.

To ensure the efficacy of LPIPS as a proxy for human judgements, we deploy two types of surveys on Amazon Me-

[1]For all evaluations we report the average over 4 different backbones used to calculate LPIPS including the finetuned weights released by the authors. More details in Appendix A.2

CIFAR10						
	MODEL	HUMAN 2AFC	LPIPS 2AFC	HUMAN CLUSTERING		LPIPS CLUSTERING
AT ℓ_2 $\epsilon = 1$	RESNET18	$96.00_{\pm 2.55}$	$87.25_{\pm 9.52}$	$97.48_{\pm 1.80}$		$88.13_{\pm 6.57}$
	VGG16	$38.83_{\pm 7.59}$	$4.00_{\pm 3.86}$	$55.39_{\pm 5.63}$		$46.09_{\pm 3.84}$
	INCEPTIONV3	$82.00_{\pm 8.44}$	$54.12_{\pm 19.23}$	$84.47_{\pm 6.32}$		$74.87_{\pm 6.74}$
	DENSENET121	$98.67_{\pm 0.24}$	$91.75_{\pm 8.2}$	$97.64_{\pm 2.08}$		$91.92_{\pm 6.13}$
STANDARD	RESNET18	$0.17_{\pm 0.24}$	$0.0_{\pm 0.0}$	$38.55_{\pm 1.19}$		$35.35_{\pm 3.27}$
	VGG16	$0.17_{\pm 0.24}$	$0.0_{\pm 0.0}$	$33.84_{\pm 2.70}$		$32.58_{\pm 1.04}$
	INCEPTIONV3	$0.17_{\pm 0.24}$	$0.38_{\pm 0.41}$	$38.38_{\pm 4.06}$		$36.62_{\pm 3.08}$
	DENSENET121	$9.83_{\pm 9.97}$	$0.12_{\pm 0.22}$	$42.42_{\pm 5.02}$		$37.12_{\pm 3.54}$
IMAGENET						
	MODEL	HUMAN 2AFC	LPIPS 2AFC	HUMAN CLUSTERING	HUMAN CLUSTERING HARD	LPIPS CLUSTERING
AT ℓ_2 $\epsilon = 3$	RESNET18	$93.17_{\pm 5.95}$	$53.37_{\pm 20.19}$	$96.00_{\pm 3.59}$	$87.75_{\pm 7.60}$	$65.28_{\pm 10.58}$
	RESNET50	$99.50_{\pm 0.00}$	$53.63_{\pm 20.64}$	$99.49_{\pm 0.71}$	$97.06_{\pm 3.47}$	$71.21_{\pm 9.93}$
	VGG16	$95.50_{\pm 2.12}$	$59.38_{\pm 21.48}$	$91.75_{\pm 5.22}$	$90.69_{\pm 3.13}$	$70.33_{\pm 9.78}$
STANDARD	RESNET18	$0.00_{\pm 0.00}$	$1.12_{\pm 1.67}$	$33.33_{\pm 0.00}$	-	$34.60_{\pm 0.56}$
	RESNET50	$5.33_{\pm 7.54}$	$0.38_{\pm 0.41}$	$38.38_{\pm 2.53}$	-	$35.35_{\pm 0.62}$
	VGG16	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$33.96_{\pm 2.00}$	-	$34.47_{\pm 1.49}$

Table 1: [CIFAR10 and ImageNet Surveys To Confirm Efficacy of LPIPS] We use LPIPS to simulate a human in both 2AFC and Clustering setups described in Section 2.1 and compare it with AMT worker's responses. We see that LPIPS and humans rank models similarly, thus showing that LPIPS is a reliable proxy for judging perceptual similarity of IRIs.

chanical Turk (AMT) to also elicit human similarity judgements. Prompts for these surveys are shown in Fig. 2. We received approval from the Ethical Review Board of our institute for this survey. Each survey consists of 100 images plus some attention checks to ensure the validity of our responses. The survey was estimated to take 30 minutes (even though on average our annotators took less than 20 minutes), and we paid each worker 7.5 USD per survey.

Clustering In this setting, we ask humans to match the IRIs (x_{r_i}) on the row to the most perceptually similar image (x_t) on the column (each row can only be matched to one column). A prompt for this type of a task is shown in Fig. 2b & 2c. With these responses, we calculate a quantitative measure of alignment by measuring the fraction of x_{r_i} that were correctly matched to their respective x_t. For ImageNet, we observed that a random draw of three images (*e.g.*, Fig. 2c) can often be easy to match to based on how different the drawn images (x_t) are. Thus, we additionally construct a "hard" version of this task by ensuring that the three images are very "similar" (as shown in Fig. 2b). We leverage human annotations of ImageNet-HSJ (Roads and Love 2021) to draw these similar images. More details can be found in Appendix A.

2AFC This is the exact test used to generate $\mathcal{A}$. In this setting we show the annotator a reconstructed image (x_r) and ask them to match it to one of the two images shown in the options. The images shown in the options are the seed (x_0, *i.e.*, starting value of x in Eq. 2) and the original image (x_t). Since x_r and x_t are IRIs for the model (by construction), alignment would imply humans also perceive x_r and x_t similarly. See Fig. 2a for an example of this type of survey.

2.2 Generating IRIs

Even if we assume a finite sampled set $X \sim \mathcal{D}$ (discussed in Section 2.3), there can be many samples in $\mathcal{S}$ due to the highly non-linear nature of DNNs. However, we draw on the insight that there is often some structure to the set of IRIs, that is heavily dependent on the IRI generation process. Prior work on understanding shared invariance between DNNs and humans has used representation inversion (Mahendran and Vedaldi 2014) to generate IRIs. However, IRIs generated this way depend heavily on the loss function used in representation inversion. Fig. 1 shows how different loss functions can lead to very different looking IRIs. We group these losses previously used in the literature to generate IRIs into two broad types: ***regularizer-free*** (used by (Engstrom et al. 2019b; Feather et al. 2019)), and ***human-leaning*** (used by (Olah et al. 2020; Mordvintsev, Olah, and Tyka 2015; Nguyen, Yosinski, and Clune 2015)). We also explore a third kind of ***adversarial*** regularizer, that aims to generate *controversial stimuli* (Golan, Raju, and Kriegeskorte 2020) between a DNN and a human.

Representation inversion is the task of starting with a random seed image x_0 to reconstruct a given image $x_t \in X$ from its representation $g(x_t)$ where $g(\cdot)$ is the trained DNN. The reconstructed image (x_r) is same as x_t from the DNN's point of view, *i.e.*, $g(x_t) \approx g(x_r)$. This is achieved by performing gradient descent on x_0 (in our experiments we use SGD with a learning rate of 0.1) to minimize a loss of the following general form:

$$\mathcal{L}_x = \frac{||g(x_t) - g(x)||_2}{||g(x_t)||_2} + \lambda * \mathcal{R}(x) \tag{2}$$

where λ is an appropriate scaling constant for regularizer $\mathcal{R}$. All of these reconstructions induce representations in the DNN that are very similar to the given image (x_t), as

measured using ℓ_2 norm. Depending on the choice of seed x_0 and the choice of $\mathcal{R}$, we get different reconstructions of x_t thus giving us a set of inputs $\{x_t, x_{r_1}, ..., x_{r_k}\}$ that are all mapped to similar representations by $g(\cdot)$. Doing this for all $x_t \in X$, we get the IRIs , $\mathcal{S} = \{X, X^{r_1}, ..., X^{r_k}\}$.

In practice we find that the seed x_0 does not have any significant impact on the measurement of shared invariance. However, the choice of $\mathcal{R}$ *does* significantly impact the invariance measurement (as also noted by (Olah, Mordvintsev, and Schubert 2017)). We identify the following distinct categories of IRIs based on the choice of $\mathcal{R}$.

Regularizer-free. These methods do not use a regularizer, *i.e.*, $\mathcal{R}(x) = 0$.

human-leaning regularizer. This kind of a regularizer purposefully puts constraints on x such that the reconstruction has some "meaningful" features. A widely used regularizer is $\mathcal{R}(x) = TV(x) + ||x||_p$ where TV is the total variation in the image. Intuitively this penalizes high frequency features and smoothens the image to make it look more like natural images. Other works achieve a similar kind of high frequency penalization by blurring x before each optimization step. We combine both these frequency-based regularizers with preconditioning in the Fourier domain and robustness to small transformations. More details can be found in Appendix A.4. Intuitively a regularizer from this category generates IRIs that have been "biased" to look meaningful to humans.

Adversarial regularizer. We propose a new regularizer to generate IRIs while intentionally making them look *perceptually dissimilar* from the target, *i.e.*, $\mathcal{R} = -||g_{\text{human}}(x_t) - g_{\text{human}}(x)||$ (negative sign since we want to maximize perceptual distance between x and x_t). We leverage LPIPS (Learned Perceptual Image Patch Similarity), a widely used *perceptual distance* measure, to approximate $||g_{\text{human}}(x_t) - g_{\text{human}}(x)||$. LPIPS uses initial layers of an ImageNet trained model (fine-tuned on a dataset of human similarity judgements) to approximate perceptual distance between images which makes it differentiable and thus can be easily plugged into Eq. 2. Thus, the regularizer used is $\mathcal{R}(x) = -\text{LPIPS}(x, x_t)$. IRIs generated using this regularizer can be thought of as *controversial stimuli* (Golan, Raju, and Kriegeskorte 2020) – they're similar from the DNN's perspective, but distinct from a human's perspective.

2.3 Choice of Inputs X

We try out many different distributions, including the training data distribution and random noise distributions, and find that takeaways about a alignment of model's invariances with humans *do not* depend heavily on the choice of X. More discussion and results in Appendix A.3.

2.4 Evaluation and Takeaways

For each model, we randomly picked 100 images from the data distribution along with a seed image with random pixel values. For each of the 100 images, we do representation inversion using one regularizer each from *regularizer-free*, *human-leaning*, and *adversarial*.

Reliability of using LPIPS Table 1 shows the results for the surveys conducted with AMT workers [2]. Each survey was completed by 3 workers. For a well aligned model, the scores under `2AFC` and `Clustering` should be close to 1, while for a non-aligned model scores under `2AFC` should be close to 0, and scores under `Clustering` should be close to a random guess (*i.e.*, about 33%). We see that LPIPS (with different backbone nets, e.g., AlexNet, VGG) orders models similar to human annotators for both the survey setups, thus showing that it's a reliable proxy.

Reliability of Human Annotators In Table 1, we make three major observations: 1) variance between different annotators is very low; 2) scores under `Human 2AFC` and `Human Clustering` order different models similarly; and finally, 3) even though accuracy drops for the "hard" version of ImageNet task, the relative ordering of models remains the same. These observations indicate that alignment can be reliably measured by generating IRIs and does not depend on bias in annotators. Note that AMT experiments were only performed on IRIs generated using the regularizer-free loss in Eq 2.

Impact of regularizer Table 2 shows the results of Alignment (Eq 1) for different regularizers for IRI generation. We evaluated multiple architectures of both standard and adversarially trained CIFAR10 and ImageNet models. We find that under different types of regularizers, the alignment of models can look very different. We also see that adversarial regularizer makes aligment bad for almost all models, thus showing that for the worst pick of IRIs the alignment between learned invariances and human invariances has a lot of room for improvement. Conversely, the human-leaning regularizer overestimates the alignment.

Impact of X In the case of OOD targets (x_t) we see that humans are still able to faithfully judge similarity, yielding the same ranking of models as in-distribution targets. Some results for human judgements about similarity of IRIs for out of distribution samples are shown in Table 4, Appendix A.3. As seen in Fig 4 (Appendix A.3), human-leaning regularizer does not work well for reconstructing noisy targets. This is because such regularizers explicitly remove high-frequency features from reconstructions (Olah, Mordvintsev, and Schubert 2017) and thus struggle to meaningfully reconstruct targets that contain high-frequency features. Hence, all results in Table 4, Appendix A.3 are reported on IRIs generated using regularizer-free loss.

Impact of x_0 We repeat some of the experiments with other starting points for Eq 2 and find that results are generally not sensitive to the choice of x_0. Results are included in Appendix A.5.

3 What Contributes to Learning Invariances Aligned with Humans

In recent years there have been efforts to understand how invariances in representations learnt by such networks align with those of humans (Geirhos et al. 2018; Hermann, Chen, and Kornblith 2020; Feather et al. 2019). However, how individual components of the deep learning pipeline affect the invariances learned is still not well understood. Prior works

[2] This was conducted only using IRIs from regularizer-free inversion.

CIFAR10						
Training	Model	Alignment			Clean Acc.	Robust Acc.
		Reg.-Free	Human-Aligned	Adversarial		
AT $\ell_2, \epsilon = 1$	ResNet18	$63.25_{\pm 26.23}$	$79.00_{\pm 21.94}$	$0.33_{\pm 0.47}$	80.77	50.92
	VGG16	$0.25_{\pm 0.43}$	$41.41_{\pm 16.74}$	$1.00_{\pm 1.41}$	79.84	48.36
	InceptionV3	$23.25_{\pm 25.56}$	$64.75_{\pm 24.17}$	$3.00_{\pm 4.24}$	81.57	51.02
	Densenet121	$82.75_{\pm 20.07}$	$86.25_{\pm 14.50}$	$1.33_{\pm 1.89}$	83.22	52.86
Standard	ResNet18	$0.00_{\pm 0.00}$	$21.09_{\pm 13.51}$	$1.33_{\pm 1.89}$	94.94	0.00
	VGG16	$0.00_{\pm 0.00}$	$21.88_{\pm 14.82}$	$0.00_{\pm 0.00}$	93.63	0.00
	InceptionV3	$0.00_{\pm 0.00}$	$21.88_{\pm 17.54}$	$0.33_{\pm 0.47}$	94.59	0.00
	Densenet121	$0.00_{\pm 0.00}$	$26.56_{\pm 16.90}$	$0.00_{\pm 0.00}$	95.30	0.00
IMAGENET						
Training	Model	Alignment			Clean Acc.	Robust Acc.
		Reg.-Free	Human-Aligned	Adversarial		
AT $\ell_2, \epsilon = 3$	ResNet18	$42.00_{\pm 38.33}$	$46.75_{\pm 39.37}$	$0.33_{\pm 0.47}$	53.12	31.02
	ResNet50	$51.00_{\pm 34.89}$	$45.75_{\pm 37.39}$	$14.00_{\pm 3.74}$	62.83	38.84
	VGG16	$55.50_{\pm 34.14}$	$55.50_{\pm 38.29}$	$11.00_{\pm 3.74}$	56.79	34.46
Standard	ResNet18	$0.00_{\pm 0.00}$	$17.00_{\pm 28.30}$	$0.00_{\pm 0.00}$	69.76	0.01
	ResNet50	$0.00_{\pm 0.00}$	$16.25_{\pm 26.42}$	$0.00_{\pm 0.00}$	76.13	0.00
	VGG16	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	$0.00_{\pm 0.00}$	73.36	0.16

Table 2: [CIFAR10 and ImageNet Model Alignment Results for Different Regularizers] Ranking of models can look very different for different regularizers. Comparing Adversarially Trained (AT) Resnet18 vs InceptionV3 on CIFAR10, we see that reguarizer-free inversion leads to Resnet18 being significantly more aligned, but the trend is much less pronounced for the human-leaning regularizer.

claim that adversarially robust models tend to learn representations with a "human prior" (Kaur, Cohen, and Lipton 2019; Engstrom et al. 2019b). This leads to the question: how do other factors such as architecture, training paradigm, and data augmentation affect the invariances of representations?

We explore these questions in this section. All evaluations in this section are based on regularizer-free IRIs. We chose regularizer-free loss over the adversarial loss as the latter shows worst case alignment for all models, which is not useful for understanding the effect of various factors in the deep learning pipeline (Appendix A.4 shows more results using the adversarial regularizer). Similarly, we preferred regularizer-free over human-leaning loss as the latter has a strong 'bias' enforced by the regularizer. While our approach generalizes to any layer, unless stated otherwise, all measurements of alignment are on the penultimate layer of the network.

3.1 Architectures and Loss Functions

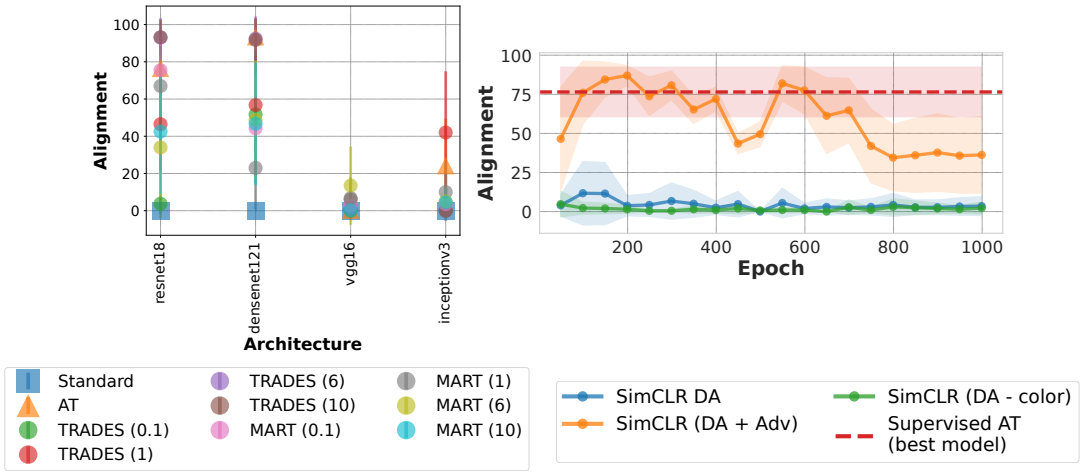
We test the alignment of different DNNs trained using various loss functions – standard cross-entropy loss, adversarial training (AT), and variants of AT (TRADES (Zhang et al. 2019), MART (Wang et al. 2019)). Both TRADES and MART have two loss terms – one each for clean and adversarial samples, which are balanced via a hyperparameter β. We report results for multiple values of β in Fig 3a and find that the alignment of standard models (blue squares) is considerably worse than the robust ones (triangles and circles). However, the effect is also influenced by the choice of model architecture, *e.g.*, for CIFAR10, for all robust training losses, VGG16 has significantly lower alignment than other architectures.

3.2 Data Augmentation

If adversarial training – which augments adversarial samples during training – generally leads to better aligned representations, then how do hand-crafted data augmentations affect invariances of learned representations? For adversarially trained models, we try with and without the usual data augmentation (horizontal flip, color jitter, and rotation). Since standard models trained with usual data augmentation show poor alignment (Section 3.1), we try stronger data augmentation (a composition of random flip, color jitter, grayscale and gaussian blur, as used in SimCLR) to see if hand-crafted data augmentations can improve alignment. Table 5 Appendix C shows how hand-crafted data augmentation can be crucial in learning aligned representations for some models (*e.g.*, adversarially trained ResNet18 benefits greatly from data augmentation). In other cases data augmentation never hurts the alignment. We also see that standard models do not gain alignment even with stronger hand-crafted data augmentations. CIFAR100 and ImageNet results can be found in Table 6 Appendix C with similar takeaways.

3.3 Learning Paradigm

Since data augmentations (both adversarial and hand-crafted) along with residual architectures help alignment, self-supervised learning (SSL) models – which explicitly rely on data augmentations – should learn well aligned representations. This leads to a natural question: how do SSL models compare with the alignment of supervised models? SimCLR is a widely used contrastive SSL method that learns 'meaningful' representations without using any labels. Recent works have built on SimCLR to also include adversarial data augmentations. We train both the standard version of SimCLR



(a) VGG16 has very low levels of alignment despite being trained using robust training losses, showing that architectures play an important role in alignment.

(b) Combining SimCLR's data augmentations (DA) with adversarial augmentations (Adv) leads to best alignment (in the early and mid epochs).

Figure 3: Role of Loss Function in Alignment (left); Role of Training Paradigm in Alignment (right); ResNet18, CIFAR10

and the one with adversarial augmentation on CIFAR10 and compare their alignment with the supervised counterparts. More training details are included in Appendix C. Additionally we also train SimCLR without the color distortion transforms – which were identified as key transforms by its authors – to see how transforms that are crucial for generalization affect alignment. Fig 3b shows the results when comparing self-supervised and supervised learning. We see that SimCLR when trained with both hand-crafted and adversarial augmentations has the best alignment, even outperforming the best adversarially trained supervised model in initial and middle epochs of training. We also see that removing color based augmentations (DA - color) does not have a significant impact on alignment, thus showing that certain DA can be crucial for generalization but not necessarily for alignment.

Summary We find that there are three key components that lead to good alignment: architectures with residual connections, adversarial data augmentation using ℓ_2 threat model, and a (self-supervised) contrastive loss. We leave a more comprehensive study of the effects of these training parameters on alignment for future work.

4 Related Work

Robust Models Several methods have been introduced to make deep learning models robust against adversarial attacks (Papernot et al. 2016; Ross and Doshi-Velez 2018; Gu and Rigazio 2014; Tramèr et al. 2018; Cohen, Rosenfeld, and Kolter 2019). These works try to model a certain type of human invariance (small change to input that does not change human perception) and make the model also learn such an invariance. Our work, on the other hand, aims to evaluate what invariances have already been learned by a model and how they align with human perception. **DNNs and Human Perception** Neural networks have been used to model many perceptual properties such as quality (Amirshahi, Pedersen, and Yu 2016; Gao et al. 2017) and closeness (Zhang et al. 2018) in the image space. Recently there has been interest in measuring the alignment of human and neural network perception. Roads et al. do this by eliciting similarity judgements from humans on ImageNet inputs and comparing it with the outputs of neural nets (Roads and Love 2021). Our work, however, explores alignment in the opposite direction, *i.e.*, we measure if inputs that a network seed the same are also the same for humans. (Feather et al. 2019; Engstrom et al. 2019b) are closest to our work as they also evaluate alignment from model to humans, however as discussed in Section 2, unlike our work, their approaches are not scalable, they do not discuss the effects of loss function used to generate IRIs, and they do not contribute to an understanding of what components in the deep learning pipeline lead to learning human-like invariances.

5 Conclusion and Broader Impacts

Our work offers insights into how measures of alignment can vary based on different loss functions used to generate IRIs. We believe that when it is done carefully, measuring alignment is a useful model evaluation tool that provides insights beyond those offered by traditional metrics such as clean and robust accuracy, enabling better alignment of models with humans. We recognize that there are potentially worrying use cases against which we must be vigilant, such as taking advantage of alignment to advance work on deceiving humans. Human perception is complex, nuanced and discontinuous (Stankiewicz and Hummel 1996), which poses many challenges in measuring the alignment of DNNs with human perception (Guest and Love 2017). In this work, we take a step toward defining and measuring the alignment of DNNs with human perception. Our proposed method is a necessary but not sufficient condition for alignment and, thus, must be used carefully and supplemented with other checks, including domain expertise. By presenting this method, we hope for better design, understanding, and auditing of DNNs.

Acknowledgements

AW acknowledges support from a Turing AI Fellowship under grant EP/V025279/1, The Alan Turing Institute, and the Leverhulme Trust via CFI. VN, AM, CK, and KPG were supported in part by an ERC Advanced Grant "Foundations for Fair Social Computing" (no. 789373). VN and JPD were supported in part by NSF CAREER Award IIS-1846237, NSF D-ISN Award #2039862, NSF Award CCF-1852352, NIH R01 Award NLM013039-01, NIST MSE Award #20126334, DARPA GARD #HR00112020007, DoD WHS Award #HQ003420F0035, ARPA-E Award #4334192 and a Google Faculty Research Award. BCL was supported by Wellcome Trust Investigator Award WT106931MA and Royal Society Wolfson Fellowship 183029. All authors would like to thank Nina Grgić-Hlača for help with setting up AMT surveys.

References

Amirshahi, S. A.; Pedersen, M.; and Yu, S. X. 2016. Image quality assessment by comparing CNN features between images. *Journal of Imaging Science and Technology*, 60(6): 60410–1.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; and Wang, Z. 2020b. Adversarial robustness: From self-supervised pretraining to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 699–708.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, 1310–1320. PMLR.

Engstrom, L.; Ilyas, A.; Salman, H.; Santurkar, S.; and Tsipras, D. 2019a. Robustness (Python Library). https://github.com/MadryLab/robustness.

Engstrom, L.; Ilyas, A.; Santurkar, S.; Tsipras, D.; Tran, B.; and Madry, A. 2019b. Adversarial Robustness as a Prior for Learned Representations. arXiv:1906.00945.

Falcon, W. 2019. PyTorch Lightning. https://github.com/PyTorchLightning/pytorch-lightning.

Feather, J.; Durango, A.; Gonzalez, R.; and McDermott, J. 2019. Metamers of neural networks reveal divergence from human perceptual systems. *Advances in Neural Information Processing Systems*, 32.

Fechner, G. T.; Howes, D. H.; and Boring, E. G. 1966. *Elements of psychophysics*, volume 1. Holt, Rinehart and Winston New York.

Gao, F.; Wang, Y.; Li, P.; Tan, M.; Yu, J.; and Zhu, Y. 2017. Deepsim: Deep similarity for image quality assessment. *Neurocomputing*, 257: 104–114.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2018. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv:1811.12231*.

Golan, T.; Raju, P. C.; and Kriegeskorte, N. 2020. Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, 117(47): 29330–29337.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572.

Gu, S.; and Rigazio, L. 2014. Towards deep neural network architectures robust to adversarial examples. *arXiv:1412.5068*.

Guest, O.; and Love, B. C. 2017. What the success of brain imaging implies about the neural code. *Elife*, 6: e21397.

Harris, C. R.; Millman, K. J.; Van Der Walt, S. J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N. J.; et al. 2020. Array programming with NumPy. *Nature*, 585(7825): 357–362.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hermann, K.; Chen, T.; and Kornblith, S. 2020. The origins and prevalence of texture bias in convolutional neural networks. *Advances in Neural Information Processing Systems*, 33: 19000–19015.

Hunter, J. D. 2007. Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(03): 90–95.

Kaur, S.; Cohen, J. M.; and Lipton, Z. C. 2019. Are Perceptually-Aligned Gradients a General Property of Robust Classifiers? *CoRR*, abs/1910.08640.

Kim, M.; Tack, J.; and Hwang, S. J. 2020. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33: 2983–2994.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

LeCun, Y.; Bengio, Y.; and Hinton, G. 2015. Deep learning. *nature*, 521(7553): 436–444.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2019. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:1706.06083.

Mahendran, A.; and Vedaldi, A. 2014. Understanding Deep Image Representations by Inverting Them. arXiv:1412.0035.

Mordvintsev, A.; Olah, C.; and Tyka, M. 2015. Inceptionism: Going Deeper into Neural Networks.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 427–436.

Olah, C.; Cammarata, N.; Schubert, L.; Goh, G.; Petrov, M.; and Carter, S. 2020. Zoom in: An introduction to circuits. *Distill*, 5(3): e00024–001.

Olah, C.; Mordvintsev, A.; and Schubert, L. 2017. Feature visualization. *Distill*, 2(11): e7.

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, 582–597. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Recht, B.; Roelofs, R.; Schmidt, L.; and Shankar, V. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, 5389–5400. PMLR.

Roads, B. D.; and Love, B. C. 2021. Enriching imagenet with human similarity judgments and psychological embeddings. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 3547–3557.

Ross, A.; and Doshi-Velez, F. 2018. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115: 211–252.

Salman, H.; Ilyas, A.; Engstrom, L.; Kapoor, A.; and Madry, A. 2020. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33: 3533–3545.

Santurkar, S.; Ilyas, A.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 32.

Shafahi, A.; Najibi, M.; Ghiasi, M. A.; Xu, Z.; Dickerson, J.; Studer, C.; Davis, L. S.; Taylor, G.; and Goldstein, T. 2019. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32.

Stankiewicz, B. J.; and Hummel, J. E. 1996. Categorical relations in shape perception. *Spatial vision*, 10(3): 201–236.

Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I. J.; Boneh, D.; and McDaniel, P. D. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Wang, Y.; Zou, D.; Yi, J.; Bailey, J.; Ma, X.; and Gu, Q. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*.

Wightman, R. 2019. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, 7472–7482. PMLR.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.