Joint rotational invariance and adversarial training of a dual-stream Transformer yields state of the art Brain-Score for Area V4

Anonymous Author(s)
Affiliation
Address
email

Abstract

Modern high-scoring models of vision in the brain score competition do not stem from Vision Transformers. However, in this paper, we provide evidence against the unexpected trend of Vision Transformers (ViT) being not perceptually aligned with human visual representations by showing how a dual-stream Transformer, a CrossViT *a la* Chen et al. (2021), under a joint rotationally-invariant and adversarial optimization procedure yields 2nd place in the aggregate Brain-Score 2022 competition (Schrimpf et al., 2020b) averaged across all visual categories, and at the time of the competition held 1st place for the highest explainable variance of area V4. In addition, our current Transformer-based model also achieves greater explainable variance for areas V4, IT and Behavior than a biologically-inspired CNN (ResNet50) that integrates a frontal V1-like computation module (Dapello et al., 2020). To assess the contribution of the optimization scheme with respect to the CrossViT architecture, we perform several additional experiments on differently optimized CrossViT's regarding adversarial robustness, common corruption benchmarks, mid-ventral stimuli interpretation and feature inversion. Against our initial expectations, our family of results provides tentative support for an *"All roads lead to Rome"* argument enforced via a joint optimization rule even for non biologically-motivated models of vision such as Vision Transformers.

1 Introduction

Research and design of modern deep learning and computer vision systems such as the NeoCognitron (Fukushima & Miyake, 1982), H-Max Model (Serre et al., 2005) and classical CNNs (LeCun et al., 2015) have often stemmed from breakthroughs in visual neuroscience dating from Kuffler (1953) and Hubel & Wiesel (1962). Today, research in neuroscience passes through a phase of symbiotic development where several models of artificial visual computation (mainly deep neural networks), may inform visual neuroscience (Richards et al., 2019) shedding light on puzzles of development (Lindsey et al., 2019), physiology (Dapello et al., 2020), representation (Jagadeesh & Gardner, 2022) and perception (Harrington & Deza, 2022).

Of particular recent interest is the development of Vision Transformers (Dosovitskiy et al., 2021). A model that originally generated several great breakthroughs in natural language processing (Vaswani et al., 2017), and that has now slowly begun to dominate the field of machine visual computation. However, in computer vision, we still do not understand why Vision Transformers perform so well when adapted to the visual domain (Bhojanapalli et al., 2021). Is this new excel in performance due to their self-attention mechanism; a relaxation of their weight-sharing constraint? Their greater number of parameters? Their optimization procedure? Or perhaps a combination of all these factors? Naturally, given the uncertainty of the models' *explainability*, their use has been carefully limited as a model of visual computation in biological (human) vision.

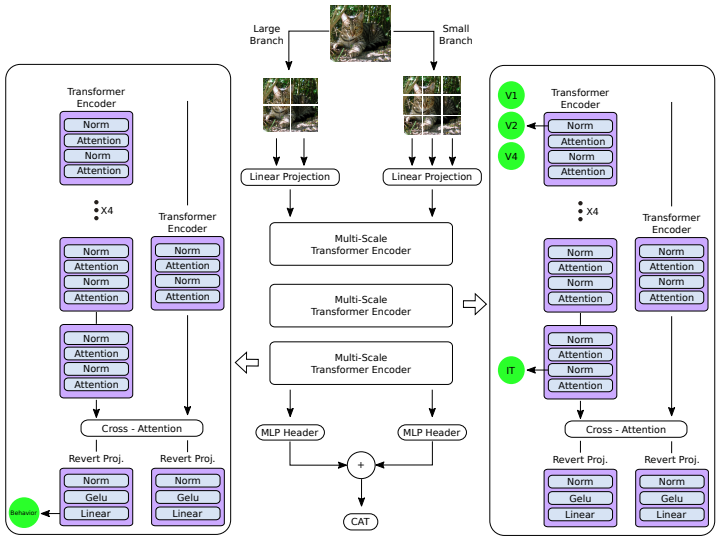


Figure 1: Diagram of CrossViT-18† (Chen et al., 2021) architecture & specification of selected layers for the V1, V2, V4, IT brain areas and the behavioral benchmark. Our Brain-Score 2022 competition entry was a variation of this model where the architecture is cloned, and the network is adversarially trained with hard data-augmentation rotations starting from a pre-trained ImageNet model.

This is a double-edged sword: On one hand, perceptual psychologists still rely heavily on relatively low-scoring ImageNet-based accuracy models such as AlexNet, ResNet & VGG despite their *limited* degree of biological plausibility (though some operations are preserved, *eg.* local filtering, half-wave rectification, pooling). On the other hand, a new breed of models such as Vision Transformers has surged, but their somewhat non-biologically inspired computations have no straightforward mapping to approximate the structure of the human ventral stream[1] – thus discarding them as serious models of the human visual system. Alas, even if computer vision scientists may want to remain on the sidelines of the usefulness of a biological/non-biological plausibility debate, the reality is that computer vision systems are still far from perfect. The existence of Adversarial examples, both artificial (Goodfellow et al., 2015; Szegedy et al., 2014) and natural (Hendrycks et al., 2021b), reflects that there is still a long way to go to close the human-machine perceptual alignment gap (Geirhos et al., 2021). Beyond the theoretical milestone of closing this gap, this will be beneficial for automated systems in radiology (Hosny et al., 2018), surveillance (Deza et al., 2019), driving (Huang & Chen, 2020), and art (Ramesh et al., 2022).

These two lines of thought bring us to an interesting question that was one of the motivations of this paper: *"Are Vision Transformers good models of the human ventral stream?"* Our approach to answer this question will rely on using the Brain-Score platform (Schrimpf et al., 2020a; BrainScore-Org, 2022) and participating in their first yearly competition with a Transformer-based model. This platform quantifies the similarity via bounded [0,1] scores of responses between a computer model and a set of non-human primates. Here the ground truth is collected via neurophysiological recordings and/or behavioral outputs when primates are performing psychophysical tasks, and the scores are computed by some derivation of Representational Similarity Analysis (Kriegeskorte et al., 2008) when pitted against artificial neural network activations of modern computer vision models.

Altogether, if we find that a specific model yields high Brain-Scores, this may suggest that such flavor of Vision Transformers-based models obey a necessary but not sufficient condition of biological plausibility – or at least relatively so with respect to their Convolutional Neural Network (CNN) counter-parts. As it turns out, we will find out that the answer to the previously posed question is complex, and depends heavily on how the artificial model is optimized (trained). Thus the main contribution of this paper is to understand *why* this particular Transformer-based model when optimized under certain conditions performs vastly better in the Brain-Score competition achieving

[1]Even at their start, the patch embedding operation is not obviously mappable to retinal, LGN, or V1-like primate computation.

			Brain-Score						ρ-Hierarchy
Rank	Model ID #	Description	Avg	V1	V2	V4	IT	Behavior	
1	1033	Bag of Tricks (Riedel, 2022) [New SOTA]	**0.515**	**0.568**	**0.360**	0.481	0.514	**0.652**	-0.2
2	991	CrossViT-18† (Adv + Rot) [Ours]	0.488	0.493	0.342	**0.514**	**0.531**	0.562	**+0.8**
3	1044	Gated Recurrence (Azeglio et al., 2022)	0.463	0.509	0.303	0.482	0.467	0.554	-0.4
4	896	N/A	0.456	0.538	0.336	0.485	0.459	0.461	-0.4
5	1031	N/A	0.453	0.539	0.332	0.475	0.510	0.410	-0.2

Table 1: Ranking of all entries in the Brain-Score 2022 competition as of February 28th, 2022. Scores in **blue** indicate **world record** (highest of all models at the time of the competition), while scores in **bold** display the highest scores of **competing entries**. Column ρ-Hierarchy indicates the Spearman rank correlation between per-Area Brain-Score and Depth of Visual Area (V1 $\rightarrow$ IT).

SOTA in such benchmark, and *not* to develop another competitive/SOTA model for ImageNet (which has shown to not be a good target Beyer et al. (2020)). The authors firmly believe that the former goal tackled in the paper is much under-explored compared to the latter, and is also of great importance to the intersection of the visual neuroscience and machine learning communities.

2 Optimizing a CrossViT for the Brain-Score Competition

Now, we discuss an interesting finding, where amidst the constant debate of the biological plausibility of Vision Transformers – which have been deemed less biologically plausible than convolutional neural networks [2], though also see Conwell et al. (2021)) –, we find that when these Transformers are optimized under certain conditions, they may achieve high explainable variance with regards to many areas in primate vision, and surprisingly the highest score to date at the time of the competition for explainable variance in area V4, that still remains a mystery in visual neuroscience (see Pasupathy et al. (2020) for a review). Our final model and highest scoring model was based on several insights:

Adversarial-Training: Work by Santurkar et al. (2019); Engstrom et al. (2019b); Dapello et al. (2020), has shown that convolutional neural networks trained adversarially[3] yield human perceptually-aligned distortions when attacked. This is an interesting finding, that perhaps extends to vision transformers, but has never been qualitatively tested before though recent works – including this one (See Figure 4) – have started to investigate in this direction (Tuli et al., 2021; Caro et al., 2020). Thus we projected that once we picked a specific vision transformer architecture, we would train it adversarially.

Multi-Resolution: Pyramid approaches (Burt & Adelson, 1987; Simoncelli & Freeman, 1995; Heeger & Bergen, 1995) have been shown to correlate highly with good models of Brain-Scores (Marques et al., 2021). We devised that our Transformer had to incorporate this type of processing either implicitly or explicitly in its architecture.

Rotation Invariance: Object identification is generally rotationally invariant (depending on the category; *e.g.* not the case for faces (Kanwisher et al., 1998)). So we implicitly trained our model to take in different rotated object samples via hard rotation-based data augmentation. This procedure is different from pioneering work of Ecker et al. (2019) which explicitly added rotation equivariance to a convolutional neural network.

Localized texture-based computation: Despite the emergence of a *global* texture-bias in object recognition when training Deep Neural Networks (Geirhos et al., 2019) – object recognition is a compositional process (Brendel & Bethge, 2019; Deza et al., 2020). Recently, works in neuroscience have also suggested that *local* texture computation is perhaps pivotal for object recognition to either create an ideal basis set from which to represent objects (Long et al., 2018; Jagadeesh & Gardner, 2022) and/or encode robust representations (Harrington & Deza, 2022).

After searching for several models in the computer vision literature that resemble a Transformer model that ticks all the boxes above, we opted for a CrossViT-18† (that includes multi-resolution + local texture-based computation) that was trained with rotation-based augmentations and also

[2] Discussed in: URL_1 URL_2

[3] Adversarial training is the process in which an image in the training distribution of a network is perturbed adversarially (*e.g.* via PGD); the perturbed image is re-labeled to its original non-perturbed class, and the network is optimized via Empirical Risk Minimization (Madry et al., 2018).

		ImageNet (↑)	Brain-Score (↑)					
Model ID #	Description	Validation Accuracy (%)	Avg	V1	V2	V4	IT	Behavior
N/A	Pixels (Baseline)	N/A	0.053	0.158	0.003	0.048	0.035	0.020
N/A	AlexNet (Baseline)	63.3	0.424	0.508	0.353	0.443	0.447	0.370
N/A	VOneResNet50-robust (SOTA)	71.7	**0.492**	**0.531**	**0.391**	0.471	0.522	0.545
991	CrossViT-18† (Adv + Rot)	73.53	0.488	0.493	0.342	**0.514**	**0.531**	**0.562**
1084	CrossViT-18† (Adv)	64.60	0.462	0.497	0.343	0.508	0.519	0.441
1095	CrossViT-18† (Rot)	79.22	0.458	0.458	0.288	0.495	0.503	0.547
1057	CrossViT-18†	**83.05**	0.442	0.473	0.274	0.478	0.484	0.500

Table 2: A list of different models submitted to the Brain-Score 2022 competition. Scores in **bold** indicate the highest performing model per column. Scores in **blue** indicate **world record** (highest of all models at the time of the competition). All CrossViT-18† entries in the table are ours.

adversarial training (See Appendix A.3 for exact training details, our *best* model also used $p = 0.25$ grayscale augmentation, though this contribution to model Brain-Score is minimal).

Results: Our best performing model #991 achieved 2nd place in the overall Brain-Score 2022 competition (Schrimpf et al., 2020b)) as shown in Table 1. At the time of submission, it holds the first place for the highest explainable variance of area V4 and the second highest score in the IT area. Our model also currently ranks 6th across all Brain-Score submitted models as shown on the main brain-score website (including those outside the competition and since the start of the platform's conception, totaling 216). A general schematic of how Brain-Scores are calculated can be seen in Figure 2.

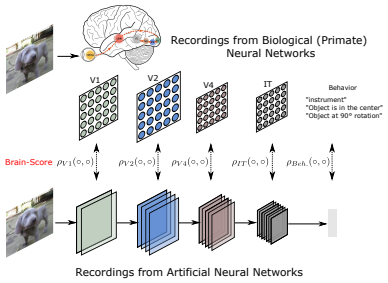


Figure 2: A schematic of how brain-score is calculated as similarity metrics obtained from neural responses and model activations.

Additionally, in comparison with the biologically-inspired model (VOneResNet50+ Adv. training), our model achieves greater scores in the IT, V4 and Behavioral benchmarks. Critically we notice that our best-performing model (#991) has a *positive* ρ-Hierarchy coefficient[4] compared to the new state of the art model (#1033) and other remaining entries, where this coefficient is negative. This was an unexpected result that we found as most biologically-driven models obtain higher Brain-Scores at the initial stages of the visual hierarchy (V1) (Dapello et al., 2020), and these scores decrease as a function of hierarchy with generally worse Brain-Scores in the final stages (*e.g.* IT).

We also investigated the differential effects of rotation invariance and adversarial training used on top of a pretrained CrossViT-18† as shown in Table 2. We observed that each step independently helps to improve the overall Brain-Score, quite ironically at the expense of ImageNet Validation accuracy (Zhang et al., 2019). Interestingly, when both methods are combined (Adversarial training and rotation invariance), the model outperforms the baseline behavioral score by a large margin (+0.062), the IT score by (+0.047), the V4 score by (+0.036), the V2 score by (+0.068), and the V1 score by (+0.020). Finally, our best model also retains a great standard accuracy at ImageNet from its pretrained version albeit a 10% drop, yet the performance on ImageNet Validation Accuracy of our model (73.53%) is still greater than a more biologically principled model such as the adversarially trained VOneResNet-50 (71.7%) (Dapello et al., 2020).

3 Assessment of CrossViT-18†-based models

As we have seen that the *optimization* procedure heavily influences the brain-score of each CrossViT-18† model, and thus its alignment to human vision (at a coarse level accepting the premise of the Brain-Score competition). We will now explore how different variations of such CrossViT's change as a function of their training procedure, and thus their learned representations via a suite of experiments

[4] ρ-Hierarchy coefficient: We define this as the Spearman rank correlation between the Brain-Scores of areas [V1,V2,V4,IT] with hierarchy: [1,2,3,4]

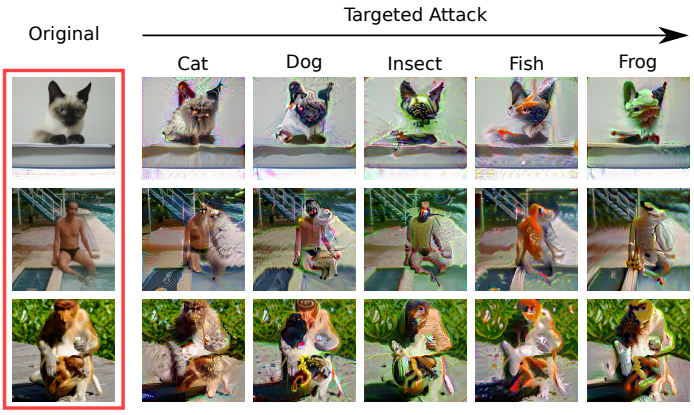


Figure 3: An extended demonstration of our winning model (CrossViT-18† [Adv. Training + Rot. invariance]) where a targeted attack is done for 3 images and the resulting stimuli is perceptually aligned with a human judgment of the fooled class. To our knowledge, this is the first time perceptually-aligned adversarially attacks have been shown to emerge in Transformer-based models.

that are more classical in computer vision. Additional experiments with CrossViT-18†-based models can be seen at Appendix B.

One of our most interesting qualitative results is that the *direction* of the adversarial attack made on our highest performing model resembles a distortion class that seems to fool a human observer too (Figures 4, 3). Alas, while the adversarial attack can be conceived as a type of *eigendistortion* as in Berardino et al. (2017) we *find* that the Brain-Score optimized Transformer models are more perceptually aligned to human observers when judging distorted stimuli. Similar results were previously found by Santurkar et al. (2019) with ResNets, though there has not been any rigorous & unlimited time verification of this phenomena in humans similar to the work of Elsayed et al. (2018). Experimental details can be found in Appendix C

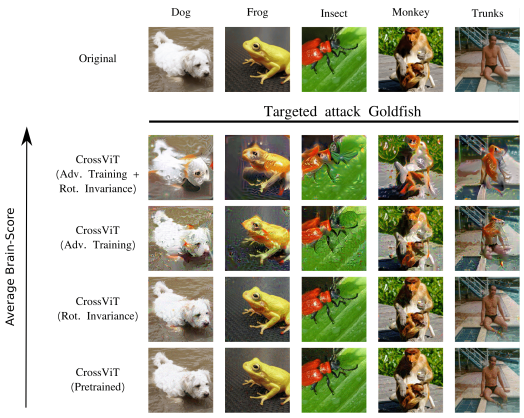
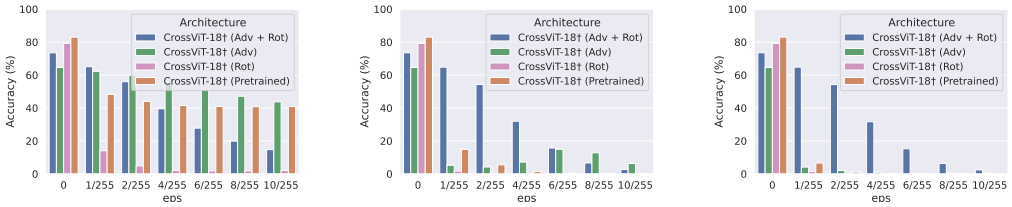


Figure 4: A qualitative demonstration of the human-machine perceptual alignment of the CrossViT-18† via the effects of adversarial perturbations. As the average Brain-Score increases in our system, the distortions seem to fool a human as well.

We also applied PGD attacks on our winning entry model (Adversarial Training + Rot. Invariance) on range $\epsilon \in \{1/255, 2/255, 4/255, 6/255, 8/255, 10/255\}$ and step-size = $\frac{2.5}{\#PGD_{iterations}}$ as in the robustness Python library (Engstrom et al., 2019a) , in addition to three other controls: Adv. Training, Rotational Invariance, and a pretrained CrossViT, to evaluate how their adversarial robustness would change as a function of this particular distortion class. When doing this evaluation we observe in Figure 5 that Adversarially trained models are more robust to PGD attacks (three-step size flavors: 1 (FGSM), 10 & 20). One may be tempted to say that this is "expected" as the adversarially trained networks would be more robust, but the type of adversarial attack on which they are trained is different (FGSM as part of FAT (Wong et al., 2020) during training; and PGD at testing). Even if FGSM can be interpreted as a 1 step PGD attack, it is not obvious that this type of generalization would occur. In fact, it is of particular interest that the Adversarially trained CrossViT-18† with "fast adversarial training" (FAT) shows greater robustness to PGD 1 step attacks when the epsilon value used at testing time is very close to the values used at training (See Figure 5a). Naturally, for PGD-based attacks where the step size is greater (10 and 20; Figs. 5b,5c), our winning entry model achieves greater robustness against all other trained CrossViT's independent of the ϵ values.



(a) PGD attack - 1 step (b) PGD attack - 10 step (c) PGD attack - 20 step

Figure 5: A suite of multiple steps [1,10,20] PGD-based adversarial attacks on clones of CrossViT-18† models that were optimized differently. Here we see that our winning entry (Adversarial training + Rotation Invariance) shows greater robustness (adversarial accuracy) than all other models as the number of steps of PGD-based attacks increases only for big step sizes of 10 & 20.

3.1 Feature Inversion

The last assessment we provided was inspired by feature inversion models that are a window to the representational soul of each model (Mahendran & Vedaldi, 2015). Oftentimes, models that are aligned with human visual perception in terms of their inductive biases and priors will show renderings that are very similar to the original image even when initialized from a noise image (Feather et al., 2019). We use the list of stimuli from Harrington & Deza (2022) to compare how several of these stimuli look like when they are rendered from the penultimate layer of a pretrained and our winning entry CrossViT-based model. A collection of synthesized images can be seen in Figure 6.

Even when these images are rendered starting from different noise images, Transformer-based models are remarkably good at recovering the structure of these images. This hints at a coherence with the results of Tuli et al. (2021) who have argued that Transformer-based models have a stronger shape bias than most CNN's (Geirhos et al., 2019). We think this is due to their initial patch-embedding stage that preserves the visual organization of the image, though further investigation is necessary to validate this conjecture.

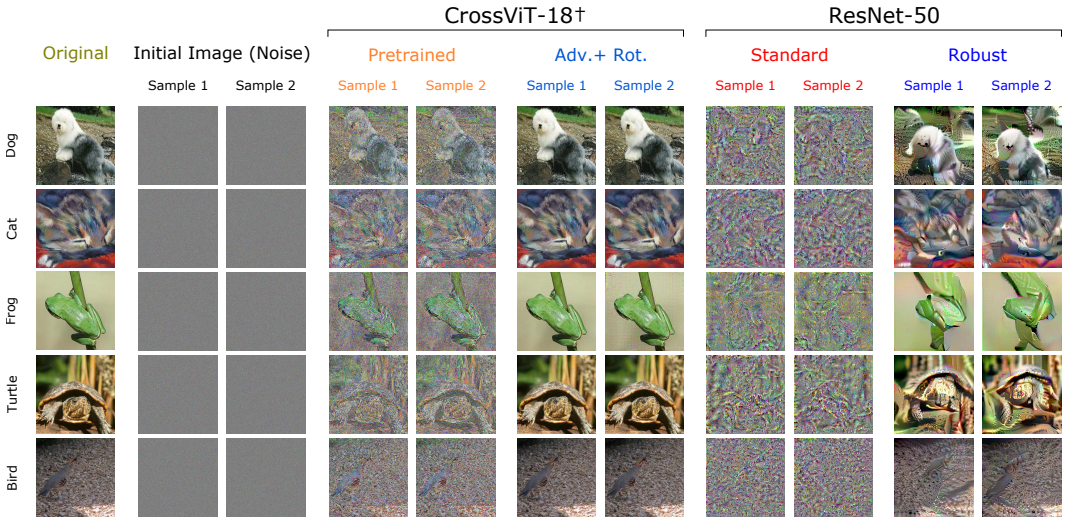


Figure 6: A summary of Feature Inversion models when applied on two different randomly samples noise images from a subset of the stimuli used in Harrington & Deza (2022). Standard and Pretrained models poorly invert the original stimuli leaving high spatial frequency artifacts.

4 Discussion

A question from this work that motivated the writing of this paper beyond the achievement of a high score in the Brain-Score competition is: How does a CrossViT-18† perform so well at explaining variance in primate area V4 without many iterations of hyper-parameter engineering? In this paper, we have only scratched the surface of this question, but some clues have emerged.

One possibility is that the cross-attention mechanism of the CrossViT-18† is a proxy for Gramian-like operations that encode local texture computation (vs global *a la* Geirhos et al. (2019)) which have been shown to be pivotal for object representation in humans (Long et al., 2018; Jagadeesh & Gardner, 2022; Harrington & Deza, 2022). This initial conjecture is corroborated by our image inversion effects (Section 3.1) where we find that CrossViT's preserves the structure stronger than Residual Networks (ResNets), while vanilla ViT's shows strong grid-like artifacts.

Equally relevant throughout this paper has been the critical finding of the role of the optimization procedure and the influence it has on achieving high Brain-Scores – even for non-biologically plausible architectures (Riedel, 2022). Indeed, the simple combination of adding rotation invariance as an implicit inductive bias through data-augmentation, and adding "worst-case scenario" (adversarial) images in the training regime seems to create a perceptually-aligned representation for neural networks (Santurkar et al., 2019).

On the other hand, the contributions to visual neuroscience from this paper are non-obvious. Traditionally, work in vision science has started from investigating phenomena in biological systems via psychophysical experiments and/or neural recordings of highly controlled stimuli in animals, to later verify their use or emergence when engineered in artificial perceptual systems. We are now in a situation where we have "by accident" stumbled upon a perceptual system that can successfully model (with half the full explained variance) visual processing in human area V4 – a region of which its functional goal still remains a mystery to neuroscientists (Vacher et al., 2020; Bashivan et al., 2019) –, giving us the chance to reverse engineer and dissect the contributions of the optimization procedure to a fixed architecture. We have done our best to pin-point a causal root to this phenomena, but we can only make an educated guess that a system with a cross-attention mechanism can *even under regular training* achieve high V4 Brain-Scores, and these are maximized when optimized with our joint adversarial training and rotation invariance procedure.

Ultimately, does this mean that Vision Transformers are good models of the Human Ventral Stream? We think that an answer to this question is a response to the nursery rhyme: *"It looks like a duck, and walks like a duck, but it's not a duck!"* One may be tempted to affirm that it is a duck if we are only to examine the family of in-distribution images from ImageNet at inference; but when out of distribution stimuli are shown to both machine and human perceptual systems we will have a chance to accurately assess their degree of perceptual similarity[5]. We can tentatively expand this argument further by studying adversarial images for both perceptual systems (See also Figure 7). Future images used in the Brain-Score competition that will better assess human-machine representational similarity should use these adversarial-like images to test if the family of mistakes that machines make are similar in nature than to the ones made by humans (See For example Golan et al. (2020)). If that is to be the case, then we are one step closer to building machines that can *see* like humans.

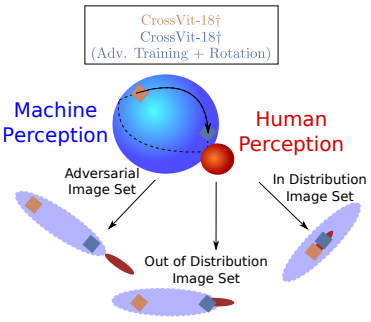


Figure 7: A cartoon inspired by Feather et al. (2019, 2021) depicting how our model changes its perceptual similarity depending on its optimization procedure. The arrows outside the spheres represent projections of such perceptual spaces that are observable by the images we show each system. While it may look like our winning model is "nearly human" it has still a long way to go, as the adversarial conditions have never been physiologically tested.

[5]Consider for example, that some stimuli used in Brain-Score are a basis set of Gabor filters, which are never encountered in nature