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

William Berrios & Arturo Deza
The Center for Brains, Minds and Machines
Massachusetts Institute of Technology
{wberrios,deza}@mit.edu

Abstract

Modern high-scoring models of vision in the brain score competition do not stem from Vision Transformers. However, in this short 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 currently (March 1st, 2022) holds the 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 Behaviour than a biologically-inspired CNN (ResNet50) that integrates a frontal V1-like computation module (Dapello et al., 2020). Our team was also the only entry in the top-5 that shows a positive rank correlation between explained variance per area and depth in the visual hierarchy. Against our initial expectations, these results provide 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. Code is available at https://github.com/williamberrios/BrainScore-Transformers

1 Optimizing a CrossViT for Brain-Score

In this short paper, we discuss an interesting finding, where amidst the constant debate of the biological plausibility of Vision Transformers – that have been deemed less biologically plausible than convolutional neural networks (as discussed in: URL_1 URL_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 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 was based on several insights:

Adversarial-Training: Work by Santurkar et al. (2019); Engstrom et al. (2019); Dapello et al. (2020), has shown that convolutional neural networks trained adversarially[1] 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 1)

[1]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).

1st Brain-Score Workshop (BSW 2022), Cascais, Portugal.

			Brain-Score						ρ-Hierarchy
Rank	Model ID #	Description	Avg	V1	V2	V4	IT	Behaviour	
1	1033	N/A [New SOTA]	**0.515**	**0.568**	**0.360**	0.481	0.514	**0.652**	-0.2
2	991	CrossViT-18†+Rotation+Adv [Ours]	0.488	0.493	0.342	**0.514**	**0.531**	0.562	**+0.8**
3	1044	N/A	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 ever-submitted to the present day), 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).

– 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 (Burt & Adelson, 1987; Simoncelli & Freeman, 1995; Heeger & Bergen, 1995) approaches 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.

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 rotation-based data augmentation. This procedure is different from pioneering work of Ecker et al. (2019) that 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 of above, we opted for a CrossViT-18† (that includes multi-resolution + local texture-based computation) that was trained with rotation-based augmentations and also 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 (Schrimpf et al., 2020b)) competition as shown in Table 1. Currently, it holds the first place for the highest explainable variance of area V4 and the second highest score in the IT area. Selected layers used from CrossViT-18† are shown in Table 2, more information can be seen in Appendix C. 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[2] 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 initial stages of the visual hierarchy (V1) (Dapello et al., 2020), and these scores decrease as a function of hierarchy with general worse Brain-Scores in the final stages (*e.g.* IT).

Table 2: Selected Layers of CrossViT-18†

Benchmark	Layer
V1,V2,V4	blocks.1.blocks.1.0.norm1
IT	blocks.1.blocks.1.4.norm2
Behavior	blocks.2.revert_projs.1.2

We also investigated the differential effects of rotation invariance and adversarial training used on top of a pretrained CrossViT-18† as shown in Table 3. We observed that each step independently

[2] 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]

		ImageNet	Brain-Score					
Model ID #	Description	Validation Accuracy (%)	Avg	V1	V2	V4	IT	Behaviour
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	voneresnet-50-robust (SOTA)	71.7	**0.492**	**0.531**	**0.391**	0.471	0.522	0.545
1057	CrossViT-18†	**83.05**	0.442	0.473	0.274	0.478	0.484	0.500
1095	CrossViT-18†+Rotation	79.22	0.458	0.458	0.288	0.495	0.503	0.547
1084	CrossViT-18†+Adv	64.60	0.462	0.497	0.343	0.508	0.519	0.441
991	CrossViT-18†+Rotation+Adv	73.53	0.488	0.493	0.342	**0.514**	**0.531**	**0.562**

Table 3: 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 ever-submitted to the present day). All CrossViT-18† entries in the table are ours.

helps to improve the overall Brain-Score. Interestingly, when both methods are combined, the model outperforms the baseline behavioral score by a large margin (+0.062). Finally, our best model also retains a great standard accuracy at ImageNet from its pretrained version.

2 Discussion

A question from this work that requires further investigation is why a CrossViT-18† performs so well at explaining variance in primate area V4 without many iterations of hyper-parameter engineering? *We do not know*, and we are currently investigating this. One possibility is that 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). However, further experiments are required to verify this hypothesis.

Finally, 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 (Figure 1). In the future we are planning on psychophysically testing this phenomenon.

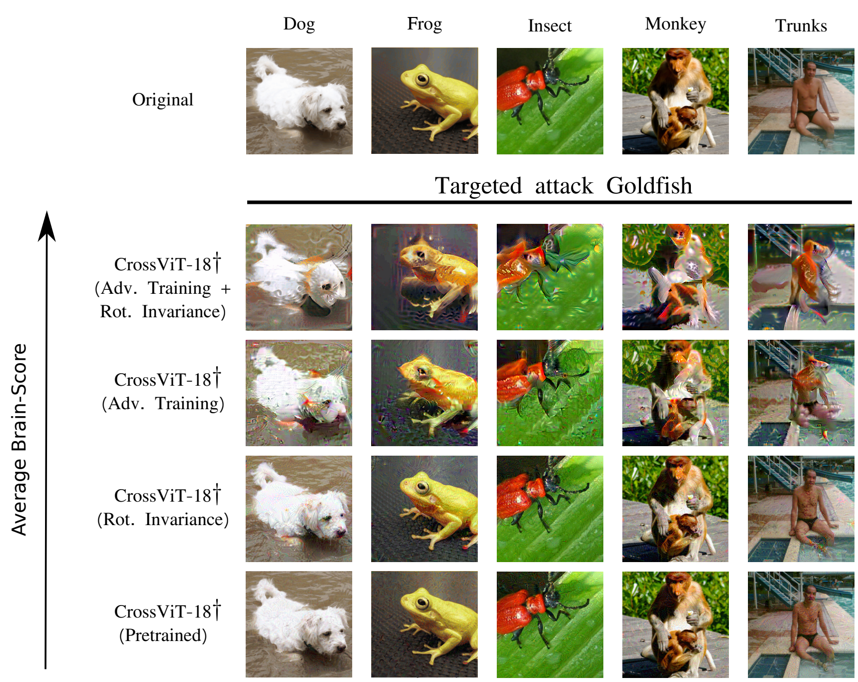

Figure 1: A qualitative demonstration of the human-machine perceptual alignment of the CrossViT-18† via the effects of adversarial perturbations. As the average Brain-Score increases in our system, the distortions seem to fool a human as well (Santurkar et al., 2019; Elsayed et al., 2018)

References

Wieland Brendel and Matthias Bethge. Approximating cnns with bag-of-local-features models works surprisingly well on imagenet. *arXiv preprint arXiv:1904.00760*, 2019.

Peter J Burt and Edward H Adelson. The laplacian pyramid as a compact image code. In *Readings in computer vision*, pp. 671–679. Elsevier, 1987.

Josue Ortega Caro, Yilong Ju, Ryan Pyle, Sourav Dey, Wieland Brendel, Fabio Anselmi, and Ankit Patel. Local convolutions cause an implicit bias towards high frequency adversarial examples. *arXiv preprint arXiv:2006.11440*, 2020.

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 357–366, 2021.

Colin Conwell, Jacob S. Prince, George A. Alvarez, and Talia Konkle. What can 5.17 billion regression fits tell us about artificial models of the human visual system? In *SVRHM 2021 Workshop @ NeurIPS*, 2021. URL `https://openreview.net/forum?id=i_xiyGq6FNT`.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James J DiCarlo. Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, 33:13073–13087, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Arturo Deza, Qianli Liao, Andrzej Banburski, and Tomaso Poggio. Hierarchically compositional tasks and deep convolutional networks. *arXiv preprint arXiv:2006.13915*, 2020.

Alexander S. Ecker, Fabian H. Sinz, Emmanouil Froudarakis, Paul G. Fahey, Santiago A. Cadena, Edgar Y. Walker, Erick Cobos, Jacob Reimer, Andreas S. Tolias, and Matthias Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=H1fU8iAqKX`.

Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=Bygh9j09KX`.

Anne Harrington and Arturo Deza. Finding biological plausibility for adversarially robust features via metameric tasks. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=yeP_zx9vqNm`.

David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pp. 229–238, 1995.

Akshay Vivek Jagadeesh and Justin Gardner. Texture-like representation of objects in human visual cortex. *bioRxiv*, 2022.

Ahmadreza Jeddi, Mohammad Javad Shafiee, and Alexander Wong. A simple fine-tuning is all you need: Towards robust deep learning via adversarial fine-tuning, 2020.

Nancy Kanwisher, Frank Tong, and Ken Nakayama. The effect of face inversion on the human fusiform face area. *Cognition*, 68(1):B1–B11, 1998.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Bria Long, Chen-Ping Yu, and Talia Konkle. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, 115(38):E9015–E9024, 2018.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=rJzIBfZAb`.

Tiago Marques, Martin Schrimpf, and James J DiCarlo. Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. *bioRxiv*, 2021.

Anitha Pasupathy, Dina V Popovkina, and Taekjun Kim. Visual functions of primate area v4. *Annual review of vision science*, 6:363–385, 2020.

Rishi Rajalingham, Elias B. Issa, Pouya Bashivan, Kohitij Kar, Kailyn Schmidt, and James J. DiCarlo. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *bioRxiv*, 2018. doi: 10.1101/240614. URL `https://www.biorxiv.org/content/early/2018/02/12/240614`.

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. *Advances in Neural Information Processing Systems*, 32, 2019.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, pp. 407007, 2020a.

Martin Schrimpf, Jonas Kubilius, Michael J. Lee, N. Apurva Ratan Murty, Robert Ajemian, and James J. DiCarlo. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, 108(3):413–423, 2020b. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2020.07.040. URL `https://www.sciencedirect.com/science/article/pii/S089662732030605X`.

Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Proceedings., International Conference on Image Processing*, volume 3, pp. 444–447. IEEE, 1995.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training, 2020.

A Experimental Setup

A.1 Dataset

We used the ImageNet 1k (Deng et al., 2009) dataset for training. ImageNet1K contains 1,000 classes and the number of training and validation images are 1.28 millions and 50,000, respectively. We validate the effectiveness of our models in the different datasets proposed in the Brain-Score (Schrimpf et al., 2020a) competition.

A.2 Custom Scheduler

The proposed learning rate scheduler is based on Jeddi et al. (2020) and is formulated as $LR = 0.00012 \times e - 0.0004$ for $e = 1$ and $LR = \frac{0.00002}{2^{e-2}}$ for $1 < e <= 6$. As shown in Figure 2, we start with a small learning rate and then it is smoothly increased for one epoch. We empirically found that fine-tuning the transformer for more than 1 epoch resulted in an under-fitting behavior of the adversarial robustness. After this first epoch, the learning rate is reduced very fast so that model performance converges to a steady state, without having too much time to overfit on the training data.

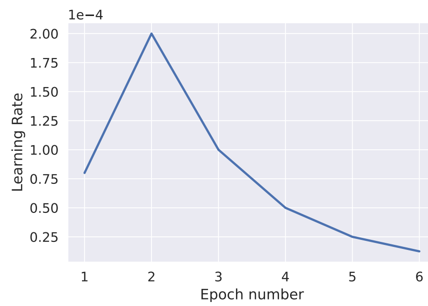

Figure 2: Custom scheduler used for training the Vision Transformer.

A.3 Training Setup

We used a pretrained CrossViT-18† (Chen et al., 2021) downloaded from the timm library that is adversarially trained via a fast gradient sign method (FGSM) attack and random initialization (Wong et al., 2020). We opted for this strategy, known as "Fast Adversarial Training" as it allows a faster iteration in comparison with other common approaches (*e.g.* adversarial training with the PGD attack). In particular, all experiments used $\epsilon = 2/255$ and step size $\alpha = 1.25\epsilon$ as proposed originally in (Wong et al., 2020). However, in contrast to the previous method, we follow a 5 epoch fine-tuning approach with a custom learning rate scheduler in order to avoid underfitting. We optimize our networks with Adaptive Moment Estimation (Adam *a la* Kingma & Ba (2014)) and employed mixed precision for faster training. All input images were pre-processed with resizing to 256×256 followed by standard random cropping and horizontal mirroring. In case of our best performing model (#991), we additionally incorporated a random grayscale transformation ($p = 0.25$) and a set of hard rotation transformations of (0°, 90°, 180°, 270°) – implicitly aiding for rotational invariance – due to the characteristics of images appearing in the behavioral benchmark of Rajalingham et al. (2018).

B Targeted Attacks in Figure 1

Table 4: Parameters used for the Goldfish targeted attack

Dataset	ϵ	Steps	Step size
ImageNet	300	500	1

C Layers Selected for Brain Areas and Behavioral Benchmark

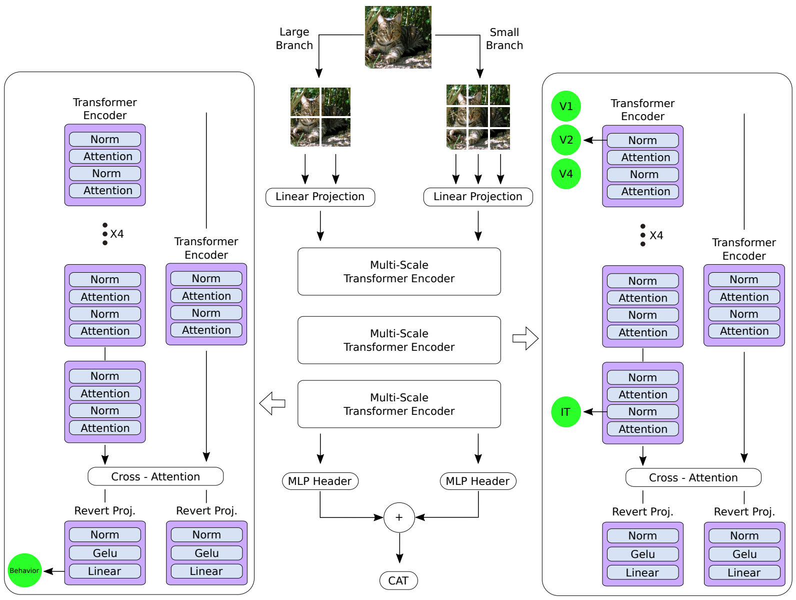

Figure 3: Diagram of CrossViT-18† (Chen et al., 2021) architecture and specification of selected layers for the V1, V2, V4, IT brain areas and the behavioral benchmark.