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 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.
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 $\rho$-Hierarchy indicates the Spearman rank correlation between per-Area Brain-Score and Depth of Visual Area (V1 → IT).

<table>
<thead>
<tr>
<th>Rank</th>
<th>Model ID #</th>
<th>Description</th>
<th>Avg</th>
<th>V1</th>
<th>V2</th>
<th>V4</th>
<th>IT</th>
<th>Behaviour</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1033</td>
<td>N/A [New SOTA]</td>
<td>0.515</td>
<td>0.568</td>
<td>0.360</td>
<td>0.481</td>
<td>0.514</td>
<td>0.652</td>
</tr>
<tr>
<td>2</td>
<td>991</td>
<td>CrossViT-18†+Rotation+Adv [Ours]</td>
<td>0.488</td>
<td>0.493</td>
<td>0.342</td>
<td>0.514</td>
<td>0.531</td>
<td>0.562</td>
</tr>
<tr>
<td>3</td>
<td>1044</td>
<td>N/A</td>
<td>0.463</td>
<td>0.509</td>
<td>0.303</td>
<td>0.482</td>
<td>0.467</td>
<td>0.554</td>
</tr>
<tr>
<td>4</td>
<td>896</td>
<td>N/A</td>
<td>0.456</td>
<td>0.538</td>
<td>0.336</td>
<td>0.485</td>
<td>0.459</td>
<td>0.461</td>
</tr>
<tr>
<td>5</td>
<td>1031</td>
<td>N/A</td>
<td>0.453</td>
<td>0.539</td>
<td>0.332</td>
<td>0.475</td>
<td>0.510</td>
<td>0.410</td>
</tr>
</tbody>
</table>

Table 2: Selected Layers of CrossViT-18†

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Layer</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1,V2,V4</td>
<td>blocks.1.blocks.1.0.norm1</td>
</tr>
<tr>
<td>IT</td>
<td>blocks.1.blocks.1.4.norm2</td>
</tr>
<tr>
<td>Behavior</td>
<td>blocks.2.revert_projs.1.2</td>
</tr>
</tbody>
</table>

$\rho$-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]
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.

---

**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.

<table>
<thead>
<tr>
<th>Model ID #</th>
<th>Description</th>
<th>Validation Accuracy (%)</th>
<th>ImageNet</th>
<th>Brain-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Avg</td>
<td>V1</td>
<td>V2</td>
</tr>
<tr>
<td>N/A</td>
<td>Pixels (Baseline)</td>
<td>N/A</td>
<td>0.053</td>
<td>0.158</td>
</tr>
<tr>
<td>N/A</td>
<td>AlexNet (Baseline)</td>
<td>63.3</td>
<td>0.424</td>
<td>0.508</td>
</tr>
<tr>
<td>N/A</td>
<td>voneresnet-50-robust (SOTA)</td>
<td>71.7</td>
<td><strong>0.531</strong></td>
<td>0.391</td>
</tr>
<tr>
<td>1057</td>
<td>CrossViT-18†</td>
<td><strong>83.05</strong></td>
<td>0.442</td>
<td>0.473</td>
</tr>
<tr>
<td>1095</td>
<td>CrossViT-18†+Rotation</td>
<td>79.22</td>
<td>0.458</td>
<td>0.458</td>
</tr>
<tr>
<td>1084</td>
<td>CrossViT-18†+Adv</td>
<td>64.60</td>
<td>0.462</td>
<td>0.497</td>
</tr>
<tr>
<td>991</td>
<td>CrossViT-18†+Rotation+Adv</td>
<td>73.53</td>
<td>0.488</td>
<td>0.493</td>
</tr>
</tbody>
</table>

*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


A Experimental Setup

A.1 Dataset

We used the ImageNet 1k 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 competition.

A.2 Custom Scheduler

The proposed learning rate scheduler is based on Jeddi et al. (2020) and is formulated as 
\[ LR = 0.00012 \times e^{-0.0004} \text{ for } e = 1 \text{ and } \\
LR = \frac{0.00002}{2^{-e}} \text{ for } 1 < e <= 6. \]
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.

A.3 Training Setup

We used a pretrained CrossViT-18 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 overfitting. 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 x 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( \epsilon )</th>
<th>Steps</th>
<th>Step size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ImageNet</td>
<td>300</td>
<td>500</td>
<td>1</td>
</tr>
</tbody>
</table>
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.