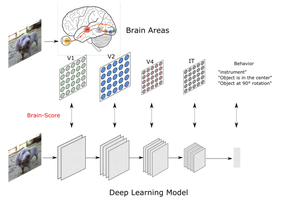
Joint rotational invariance and adversarial training of a dual-stream Transformer yields state of the art Brain-Score for area V4

William Berrios & Arturo Deza

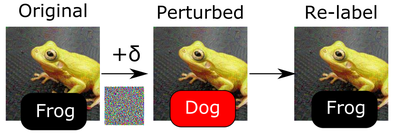
Can Vision Transformer be perceptually aligned to human visual representations?

Despite the current trend of Vision Transformers (ViT) being not perceptuallly aligned with human visual representations, we demostrate that a CrossViT under a joint rotationally invariant and adversarial optimization yields 2nd place at Brain-Score Competition and held the 1st place for the highest explainable variance in V4 area at the time of the competition.



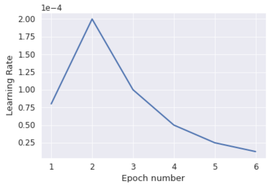
How to optimize a CrossViT to be more brain-like?

* Adversarial Training



* Hard Rotation Augmentation

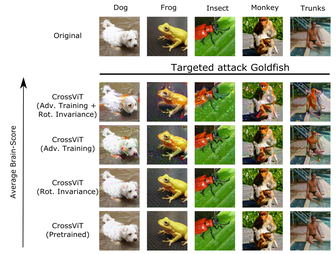
* Custom Scheduler

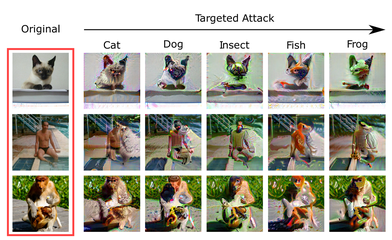


Qualitive Assesments of CrossViT Models

Human Visual Aligment against Targeted Attack

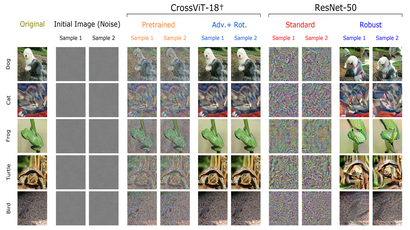
As the average Brain-Score increases in our system, the distortions seem to fool a human as well.





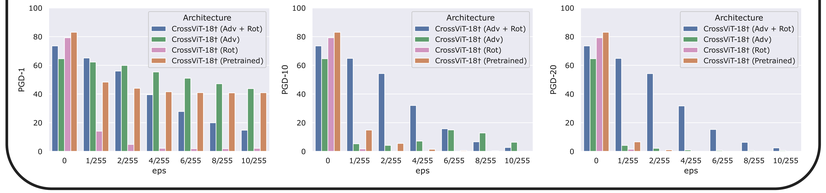
Feature Inversion

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



Adversarial Robustness

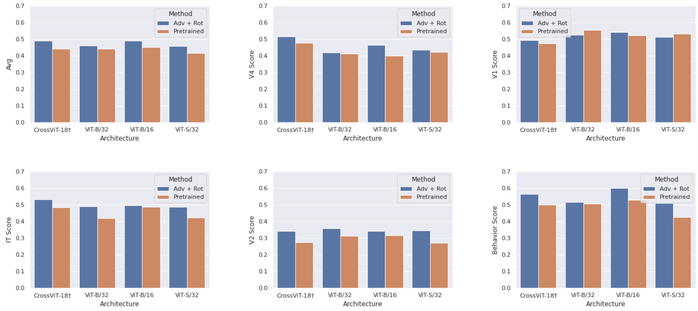
Naturally, for PGD-based attacks where the step size is greater (10 and 20), our winning entry model achieves greater robustness against all other trained CrossViT's independent of the ε values.



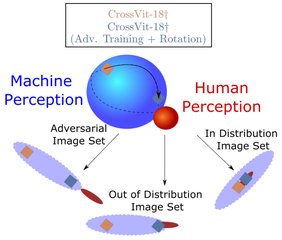
Brain Score Analysis on CrossViT & Vanilla Transformers

To our knowledge this is the first time that it has been shown that adversarial training coupled with rotational invariance homogeneously increase brain-scores across Transformer-like architectures

Brain-Score on V2, V4, superior processing IT, Behaviour and **Average** increase independent of the type of Vision Transformer used for our suite of models



Does this mean the brain is a Transformer? (or a Transformer is a Brain?)



Check our code and paper!!

A specific model that yields high Brain-Scores may suggest that such flavor of Vision Transformers-based models obey a necessary but not sufficient condition of biological plausibility.

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 perceptul systems we will have a chance to accurately as sess their degree of perceptual similarity