A Deep Learning Approach to Understanding Real-World Scene Perception in Autism

Erica L. Busch
Honors Thesis in Cognitive Science, Dartmouth College
Advisors: Caroline Robertson (Dartmouth College Psychological and Brain Sciences) and Leyla Isik (Johns Hopkins Cognitive Science)
Awarded High Honors in March 2020

Abstract

Autism is a multifaceted neurodevelopmental condition. Around 90% of individuals with autism experience sensory sensitivities. Despite this high percentage, previous studies of visual perception in autism severely limit our understanding of this symptom due to their experimental designs. Many of these studies use highly unrealistic stimuli and experimental methods or report unreproducible, conflicting results. In this study, we use a cutting-edge approach to investigate the nature of real-world visual perception in autism. First, we use virtual reality headsets with built-in eye-trackers to measure gaze behavior while individuals freely explore real-world scenes. Then, we compare gaze behavior with the representations within convolutional neural networks (CNNs), a class of computational models resemblant of the primate visual system. This allows us to model the stages of the visual processing hierarchy that could account for differences in visual processing between individuals with and without autism. To our knowledge, this is the first fully unbiased, data-driven approach to studying naturalistic visual behavior in autism. In brief, we found that convolutional neural networks, regardless of the task upon which they were trained, are better able to predict gaze behavior in typically developing controls than in individuals with autism. This suggests that differences in gaze behavior between the two groups are not principally driven by the semantically-meaningful features within a scene and emerge from differences earlier in visual processing.

Introduction

Characteristics of autism spectrum conditions

Autism spectrum condition (henceforth ASC or autism) is a complex neurodevelopmental condition affecting one out of fifty-nine individuals in the United States (CDC, 2019). Despite autism's prevalence, we know relatively little about its underlying neurobiology. The current literature on the possible neural underpinnings of autism is rife with inconsistencies, as are the rates of diagnosis worldwide (ranging from 1 in 27 in Hong Kong to 1 in 3,333 in Poland) (Elsabbagh et al., 2012).

On average, individuals are diagnosed with autism at age four in the U.S., although signs of autism are often visible by nine months. Sensory processing differences are particularly notable early in development and predictive of later autism diagnosis (Baranek et al., 2013; Estes et al., 2015), and thus may serve as promising early markers of the condition (Robertson & Baron-Cohen, 2017). In adults, atypical sensory perception impacts over 90% of individuals with ASC (Tavassoli et al., 2014). As of 2014, sensory reactivity is included in the DSM-5 criteria for ASC. This new diagnostic criterion emphasizes the important role sensory perception and processing plays in ASC, and recent studies focus on these atypicalities as defining features of the condition's neural underpinnings (Robertson & Baron-Cohen, 2017). Such sensory atypicalities include visual, tactile, taste, gustatory, and auditory sensitivities significantly different from control populations and populations with other clinical conditions (fragile X syndrome and developmental disabilities of mixed etiology), appear before three years of age, and persist through adulthood (Kern et al., 2006; Rogers et al., 2003).

Visual perception in autism

A common characterization of perception in autism is as emphasizing local details at the expense of the global percept, i.e. 'Seeing the trees, but not the forest' (Dakin & Frith, 2005). In other words, they are exceptionally attentive to visual details rather than global features, which (in a subgroup of individuals with autism) is linked to savant-like drawing abilities (Baron-Cohen et al., 2009; Mottron & Belleville, 1993). Numerous studies have shown that individuals with autism are faster at detecting visual targets among distractors ('a tree within the forest'), and eye-tracker studies have shown this as an early behavioral marker of the condition in toddlers (Gliga et al., 2015; Kaldy et al., 2011; Plaisted et al., 1998). Yet, little is known about how such a detailed perceptual style manifests in real-world environments, and novel computational methods have rarely reached autism research.

One previous study attempted to use machine learning to characterize visual processing in autism in real-world scene images. Wang et al. (2015) showed that individuals with ASC viewing scenes on a computer screen demonstrate more pixel-level saliency than controls. They also demonstrate decreased saliency for semantic-level features, as well as faces and locations considered socially-meaningful by controls (Wang et al., 2015). For example, individuals with autism viewing naturalistic scenes show preference to regions of scenes with high contrast and color instead of regions with faces, text, or other semantically informative features (Robertson & Baron-Cohen, 2017). Comparatively low responsivity to social stimuli (and nonsocial to a lesser effect) is predictive of later autism diagnosis risk for children at as young as 11 months (Baranek et al., 2013), and decreased social attention is a hallmark of ASC behavior.

These behavioral results suggest that visual perception differences in ASC originate from differences in early visual processing. Neuroimaging studies support the conclusion that atypicalities in visual perception are linked to atypical responses in primary sensory cortices. For instance, individuals with ASC have difficulty tracking the global motion of multiple objects. This relates to atypical responses in early visual cortex (Robertson et al., 2014) and primary motion area (MT) (Herrington et al., 2007; Peiker et al., 2015; Takarae et al., 2014), as shown in numerous functional magnetic resonance imaging (fMRI) studies of motion processing. Studies have also linked motion perception abnormalities with superior temporal sulcus (STS) (Dakin & Frith, 2005), a region also implicated in perceiving and understanding social interactions (Isik et al., 2017).

Naturalistic perception

A standard measure of visual attention allocation, gaze behavior has been widely suggested as a 'behavioral marker' of autism. A key limitation of these studies, though, is how little they resemble the everyday visual experience. Most preeminent studies of vision in autism utilize stimuli like Gabors, line drawings, basic block patterns, and moving dots (Simmons et al., 2009; Spencer & O'Brien, 2006; among many others). Wang et al. (2015) used real-world scenes as stimuli to measure gaze behavior, but they were presented as still images on a computer screen. Most studies using naturalistic stimuli have focused on visual attention towards social interactions, again treating autism as a condition of the social mind primarily and sensory processing secondarily (Simmons et al., 2009).

Using naturalistic stimuli in psychological studies (neuroimaging and behavioral studies alike) generates a large mass of noisy data that researchers need to sift through to draw meaningful conclusions about cognition. Computational models are powerful tools for making sense of such data. Modern deep neural networks (DNNs) are biologically-inspired models that can solve some of the cognitive tasks once thought unique to humans. These models are built of layers of simple processing units that work in parallel and communicate in feed-forward and feed-backward projections. They are trained with massive labeled data sets to complete specific tasks, like predicting market fluctuations, natural language processing, or image recognition. This training allows a model to learn a specific set of connection weights so it can make useful predictions on unseen data (Cichy & Kaiser, 2019).

Initially built as engineering tools, DNNs and especially deep convolutional neural networks (CNNs) have received particular attention for their outstanding performance on computer vision tasks. CNNs dominated the computer vision field beginning with the AlexNet network in 2012 (Krizhevsky et al., 2012), proceeding to surpass all other computational models and even human performance at visual recognition tasks (Kriegeskorte, 2015). Beyond the

computer vision field, DNNs are remarkable models of many complex tasks including brain activity (Sak et al., 2014).

Deep neural networks and the brain

Perhaps naturally given their biological inspiration, deep neural networks have tremendously benefitted neuroscience research. DNNs predict both primate behavioral and neural responses, informing the link between the two. CNNs, known for their visual recognition prowess, strongly resemble the neural responses in primate primary sensory cortices. Khaligh-Razavi & Kriegeskorte (2014) showed that a CNN, among a pool of 37 models, best predicted inferior temporal (IT) responses to visually-presented objects. Though several models trained on low-level features predicted early visual cortex responses, later levels of the CNN model (which are notably more representative of visuo-semantic features) exceeded all other models at predicting higher-order visual areas (including FFA, PPA, and LOC). Specifically, the later layers of the CNN best predicted responses in both monkey and human IT. The network's final spatially-selective layer best predicts neural responses of all their tested models, and each layer of the model's hierarchy increases roughly monotonically in its ability to predict IT responses (Khaligh-Razavi & Kriegeskorte, 2014).

Since 2014, many studies have supported the findings in Khaligh-Razavi and Kriegeskorte: CNNs are highly predictive of visual cortex responses, with early layers better predicting early visual cortex and later layers better predicting higher-level visual areas (Cichy & Kaiser, 2019; Isik et al., 2014; Kriegeskorte, 2015). Units in early layers of CNN have small receptive fields tuned to features like edges, analogous to single neurons in the early visual cortex (Hubel & Wiesel, 1962). The activity within the early layers of a CNN is more general and less task-specific, and representations increase in task-specificity in later layers. Units further along the ventral visual pathway have larger receptive fields, are more transformationally and translationally invariant, and are more selective to particular shapes and semantic categories (Güçlü & van Gerven, 2015; Hung et al., 2005; Yamins et al., 2014), similar to IT cortex responses (Khaligh-Razavi & Kriegeskorte, 2014). This brain-model analogy has proven a useful framework: CNNs are now an integral feature of models that can predict visual stimuli from observed brain (BOLD) activity (Güçlü & van Gerven, 2015).

Convolutional neural networks, the brain, and gaze behavior

The MIT saliency benchmark presents models that best predict human gaze behavior in natural images. To date, the ten top models of the MIT saliency benchmark are all CNNs (O'Connell & Chun, 2018; MIT Saliency Benchmark). DeepGaze II, a prominent saliency estimation model, utilizes a prominent CNN architecture (VGG; Simonyan & Zisserman, 2015) trained on object recognition to predict 87% of patterns of fixation and outperforms all other models on the saliency benchmark. Though the information represented in CNN

activity depends upon the training of the network, the features a CNN learns to represent during training are transferable between tasks. Thus, CNNs learn flexible feature space for an array of objectives: one can use image recognition models to predict human gaze behavior (Kümmerer et al., 2016) and scene recognition models to predict neural responses (O'Connell & Chun, 2018).

In an elegant series of experiments, O'Connell and Chun (2018) showed a novel three-way connection between CNNs, neural responses, and gaze behavior. First, they used brain activity (BOLD responses) to directly predict gaze behavior. Then, they translated these neural responses into CNN activity patterns at specific layers of a VGG model trained to recognize scenes. These activity patterns were used to build spatial priority maps, indicating the most salient regions of a scene. Spatial priority maps were then used to reconstruct fixation maps in novel scenes and to predict human fixations. By extracting CNN representation of scenes, they linked visual cortex activity, CNN activity, and gaze behavior (O'Connell & Chun, 2018). Despite the potential for CNNs to explain real-world visual behavior, to our knowledge they have not yet been applied to 360° gaze behavior. Moreover, they have never been used to characterize visual processing in autism.

A novel approach: CNNs to understand visual processing in autism

We use convolutional neural networks to investigate gaze behavior of individuals with autism while exploring naturalistic scenes. To do this, we designed a novel experimental approach. First, we used in-headset eye-tracking in immersive virtual reality headsets to measure spatial attention allocation in real-world scenes. This affords objective insight into the day-to-day visual experiences of individuals with autism, which we compare with typically-developing controls. Second, we use CNNs to model where along the visual hierarchy scene perception and processing diverges between groups, and whether this relates to the model's training. We utilized the hierarchical representations of scenes and objects within convolutional neural networks as models of spatial attention allocation across low-level visual features (e.g. color, pixel-saliency, or contrast) and high-level ones (e.g. semantically meaningful features).

Through modeling gaze behavior with CNNs, we asked a series of questions about how individuals experience the visual world. Does visual attention allocation in natural scenes vary between ASC and typically-developing controls? Are between-group differences in visual behavior predicted by different layers along the CNN hierarchy, since such layers represent increasing semanticity? Are models trained for different visual recognition tasks predictive of group differences? This could inform how visual perception characterizes autism, and help researchers understand where along the ventral pathway differences in visual representation originate.

In brief, we found that CNNs overall predicted gaze behavior better for controls than for individuals with ASC. This group difference was not affected by location along the network hierarchy-- for all individuals, the later layers of the network better predicted gaze behavior than the earlier ones, and all layers better predicted controls than individuals with autism. We also found that a model pre-trained on object recognition significantly predicted gaze behavior better than scene recognition in both individuals with and without autism, and it better predicted control gaze behavior than ASC. This suggests that differences in gaze behavior between the two groups are not tied to high-level representation of objects or places and begin early in visual processing.

Materials and Methods

Behavioral data

Behavioral data, or 'gaze data,' was collected from participants according to the following specifications.

Participants.

Forty-one adults participated in this experiment (20 ASC). All participants were recruited from the local Upper Valley (NH/VT) community. Control participants (12 female; mean age 22.38 +/- 4.84 STD years) were included based upon 1) having normal or corrected vision and no colorblindness, 2) having no neurological or psychiatric conditions, and 3) having no history of epilepsy.

Twenty participants (8 female, 1 gender unspecified; mean age 23.4 +/- 7.19 STD years) had documented autism spectrum condition (ASC) diagnoses, confirmed with the Autism Diagnostic Observation Schedule Second Edition (ADOS-2) Module 4 assessment administered by a research-reliable administrator (Hus & Lord, 2014). ASC participants all had normal or corrected vision and no colorblindness. Fifty percent of ASC participants self-reported co-occurring conditions, including anxiety, depression, attention-deficit/hyperactivity disorder, and dyslexia. These co-occurrences were not controlled for or matched in the control group. An additional four participants with ASC attempted the experiment but were later excluded from analyses due to one of three reasons: task comprehension difficulty, insufficient diagnostic confirmation, or contributing fewer than 30 valid scenes.

All participants completed the Kaufman Brief Intelligence Test (Kaufman & Kaufman, 2014) and the Autism Spectrum Quotient (Baron-Cohen et al., 2001). Control participants were matched for age with participants with ASC. An additional eight individuals participated in a pilot experiment to identify stimuli balanced for social/nonsocial salience (for more

information, see section 3.1.2). Written consent was obtained from all participants in accordance with a protocol approved by the Dartmouth College Institutional Review Board.

Stimulus and head-mounted display.

Stimuli consisted of 360-degree "photospheres" of real-world scenes sourced from open online databases such as Flickr (flickr.com) or Youtube (youtube.com). Photospheres depicted a diverse set of indoor and outdoor settings and contained 1-3 people and non-social yet interesting objects. Pilot participant data identified a set of 60 photospheres balanced for both salient social and non-social content. Such balanced scenes were defined as ones where the top 50% of pilot participants' gaze heat is distributed across both socially meaningful features (i.e. faces or bodies) and nonsocial yet identifiable and interesting features (i.e. televisions or trees). Each photosphere was then applied to a virtual environment built in Unity version 2018.3.11f1 (unity3d.com) then integrated with a head-mounted display (Oculus Rift Development Kit 2, oculus.com, low persistence OLED screen, 2K resolution per eye, ~90-degree field of view, 75 Hz refresh).

Eye-tracking specifications.

Two in-headset binocular eye-trackers monitored participants' gaze continuously during scene viewing (Pupil Labs version 1.9.7, 120 Hz sampling frequency, 0.6 visual degrees accuracy, 0.08 visual degrees precision, 5.7ms camera latency, 3ms processing latency). Custom scripts written in C# for Unity were used to record eye movements.

Experimental procedures.

During each experimental trial, participants were presented with a photosphere via the head-mounted display for twenty seconds. Each participant had the opportunity to view all 60 photospheres. The number of trials each participant actually completed varied according to participant time restraints or fatigue. On average, control participants completed 59 trials (+/- 1.8 STD trials) whereas ASC participants completed 59 trials (+/- 3.9 STD trials). Experimental trials with insufficient or low-confidence data were excluded according to the preprocessing steps (see section 3.1.6 for pre-processing details).

During viewing, participants were told to "look around each place just like you would look around a new place in real life. Pretend like you'll have to describe that place later to someone who didn't see it." Participants were given a break after every ten scenes, at which point the eye-tracker was recalibrated. Participants stood while wearing the head-mounted display and actively explored the photosphere via self-directed eye-movements and head turns. This provided an opportunity to explore the naturalistic environment from an egocentric perspective (Figure 1).

Practice trials and calibration routine.

The experiment had three phases: practice, calibration, and experimental trials. During the practice phase, each participant saw two scenes that were not included in analyses. They were reminded to move their heads and explore the whole scene. Practice phases ensured that participants had acclimated to virtual reality environments before beginning the experiment. Then, participants performed a 21-point calibration routine (approximately one minute) to validate eye-tracking accuracy. This calibration routine was repeated after every 10 experimental trials.

After each trial of the experimental phase, participants returned to a virtual home screen where they took a five-second break before the next trial. After leaving the home screen, participants saw a pre-trial fixation screen with a visual target at the center of the screen. If significant gaze drift (>5 degrees visual angle) was detected at this time, the calibration routine was repeated. Re-calibration also occurred after every time a participant removed the headset.

Data preprocessing.

Gaze data was filtered and excluded from analysis for one of three reasons. First, we filtered for eye tracker confidence. If, for any time point in the trial, eye tracker confidence fell below 50%, we exclude that data. If we excluded more than 75% of the time points for a trial, we exclude the entire trial. Next, we filtered for adequate scene exploration. For a trial to be included in our analysis, the participant must have explored at least 60% of the scene's yaw with confident eye tracking. This ensures that the participant understood the task and actively explored the scene. Finally, we thresholded for pretrial calibration check failures. Before exploring a scene, participants fixated on a target at the scene's center and we calculated the eye tracker's drift away from their gaze so we can correct for this drift at other time points in the trial. If drift exceeded 10 degrees visual angle, we excluded that trial from analyses.

To determine fixations, we calculated the orthodromic distance and velocity between consecutive gaze points. We calculated the mean absolute deviation (MAD) in gaze position using a seven-sample sliding window of ~80ms (Voloh et al., 2019). Windows with a MAD less than 50°/s were defined as potential fixations (Peterson et al., 2016). If two group centroids were displaced by under 1° and two potential fixations occurred within 150 ms, the potential fixations were concatenated. We excluded fixations that were shorter than 100ms (Peterson et al., 2016; Wass et al., 2013) as previously defined in Haskins et al. (2020).

Figure 1: Experimental design.
Participants wore immersive virtual reality headsets equipped with binocular eye trackers. On each experimental trial, participants freely explored immersive, naturalistic environments with self-generated movements (saccades and head turns).

CNN model data.

CNN architecture.
To model spatial attention allocation, we used VGG16, a convolutional neural network (CNN) with a deep, feed-forward architecture (Simonyan & Zisserman, 2015). This architecture consists of 18 spatially selective layers broken into five blocks of convolution operations followed by non-linear max-pooling operations. Convolution layers are made of 64 to 512, 3 by 3 filters, which slide across the activation volumes in a block with a stride of 1. The max-pooling operations use a 2 by 2 filter with a stride of two. By downsampling less relevant features and propagating forward more salient ones, the network can reduce the spatial size of its representation by half and build translational invariance in its representation (Figure 2C).

CNN training.
Along this processing hierarchy, the network moves from detection of low-level features (e.g. edges and contrast) to high-level, semantically meaningful concepts. A network's 'concept' is relative to the task it was trained to complete. Thus, the later layers of a network trained on scene recognition might represent concepts like roads or churches, whereas a network trained on object recognition might represent concepts like dogs or toasters.

To investigate the differences in the tasks performed by participants upon exploring a novel scene, we modeled their gaze behavior in experimental trials using convolutional neural networks pre-trained on two different tasks: scene recognition (Places365, (Zhou et al.,

2016)) and object recognition (ImageNet, (Deng et al., 2009)). Theoretically, if individuals perform a task more similar to scene recognition in one of our experimental trials, attending to more global features of a scene, their gaze maps will resemble the top pooling-layer activation of a CNN model trained to recognize scenes. If they perform a task more focused on local details (such as the objects embedded within a scene), their gaze maps will resemble the top pooling-layer CNN map of a model trained to recognize objects. We compare models trained on two different tasks (object recognition and scene recognition) to test the local versus global attention used by individuals with and without autism when exploring novel scenes.

The hidden layer weights learned by VGG16-Places365 are available from the Places365 GitHub (*CSAILVision/places365*, 2016/2020) as a Caffe model, which we converted to Pytorch for these analyses. The hidden layer weights learned by VGG16-ImageNet are available as a pre-trained model in Pytorch (Paszke et al., 2019). Though our chosen CNNs utilize the same architecture, their training is optimized to attend to distinct visual features. VGG16-Places365 (referred to as Places-CNN) attends to the scene's global features, such as affordances and terrain, to classify a scene as one of 365 place categories. VGG16-ImageNet (referred to as Object-CNN) attends to local features embedded within a scene, such as specific items or beings, to classify an object within an image into one of 1000 categories. We use identical models with distinct hidden weight sets to quantitatively model local versus global attention during scene exploration.

Within each of our models (Places-CNN and Object-CNN), the layers along the hierarchy can be regarded as maps of the most salient features within a scene at that stage of processing. Since the models share identical architectures, we can compare feature maps from the same layer (e.g. pool-5 Places-CNN to pool-5 Object-CNN). Visualizing the activations of these feature maps as the CNN processes an image illustrates the network building its conceptual representation of an image, from low-level to high-level features. In later analyses, we capitalize upon this gradual representation construction to investigate group differences in the representation of scene features between ASC and typically-developing individuals.

Panoramic image feature map extraction.

We collect eye-tracking data from individuals freely exploring novel photospheres of real-world scenes using virtual reality headsets. The nature of this design requires spherical scenes, which need to be broken down into square 'viewport' images in order to a) account for the equirectangular distortion of projecting a spherical scene into Cartesian space, and b) feed the network an image of its preferred input size without losing spatial resolution.

We developed a novel pipeline for extracting CNN activity maps in response to a photosphere and comparing that activation to behavioral gaze data. First, we sample 500 locations on a sphere such that points are sampled with greatest density around the equator and with decreasing density as moving away from the equator. Individual gaze behavior is strongly biased toward the equator of images, with decreasing fixations at increasing latitudes above and below the equator (Judd et al., 2009; Sitzmann et al., 2016). Such sampling accounts for equatorial bias by upsampling and downsampling our photospheres accordingly (Figure 2A & 2B).

We used the Equirec2Perspec module (Fu-En.Wang, 2017/2020) to project from spherical coordinates to Cartesian coordinates and build a square viewport centered on each point. Each viewport is a 224 by 224-pixel view of the photosphere accounting for 90-degrees of visual angle, approximately equivalent to that of the Oculus Rift used in experimental trials. Each viewport is large enough to capture a meaningful portion of the image contained in one field of view while avoiding equirectangular distortion (Figure 2A). The VGG16 architecture requires input volumes of size 224 by 224 x 3 (pixels x pixels x channels), but by sampling our sphere rather than simply resizing a flattened photosphere (a panoramic image of 1000 by 2000 pixels), we feed the model an image comparable to the field of view of one fixation during an experimental trial. With 500 samples of 224 by 224 pixels, viewports overlapped with one another heavily (moreso around the equator, where we sample most densely). This allowed us to average the network's activations at a given pixel when considered in numerous contexts, thus inferring how the network responds to each pixel in the context of the whole photosphere.

After sampling the photosphere into viewports and propagating viewport volumes forward through the VGG16-Places365 and VGG16-ImageNet models, we extract each model's pooling layer activations (Figure 2C) and project that activation outward into equirectangular space using Perspec2Equirec (Fu-En.Wang, 2017/2020). We repeat this process for all 500 viewports until we have generated one CNN activity heatmap representing the activation of each desired layer in response to the whole photosphere. For the pixels in which our viewport samples overlap, within each model we average over all aggregated activation values at that pixel in order to obtain an average activation value for each pixel.

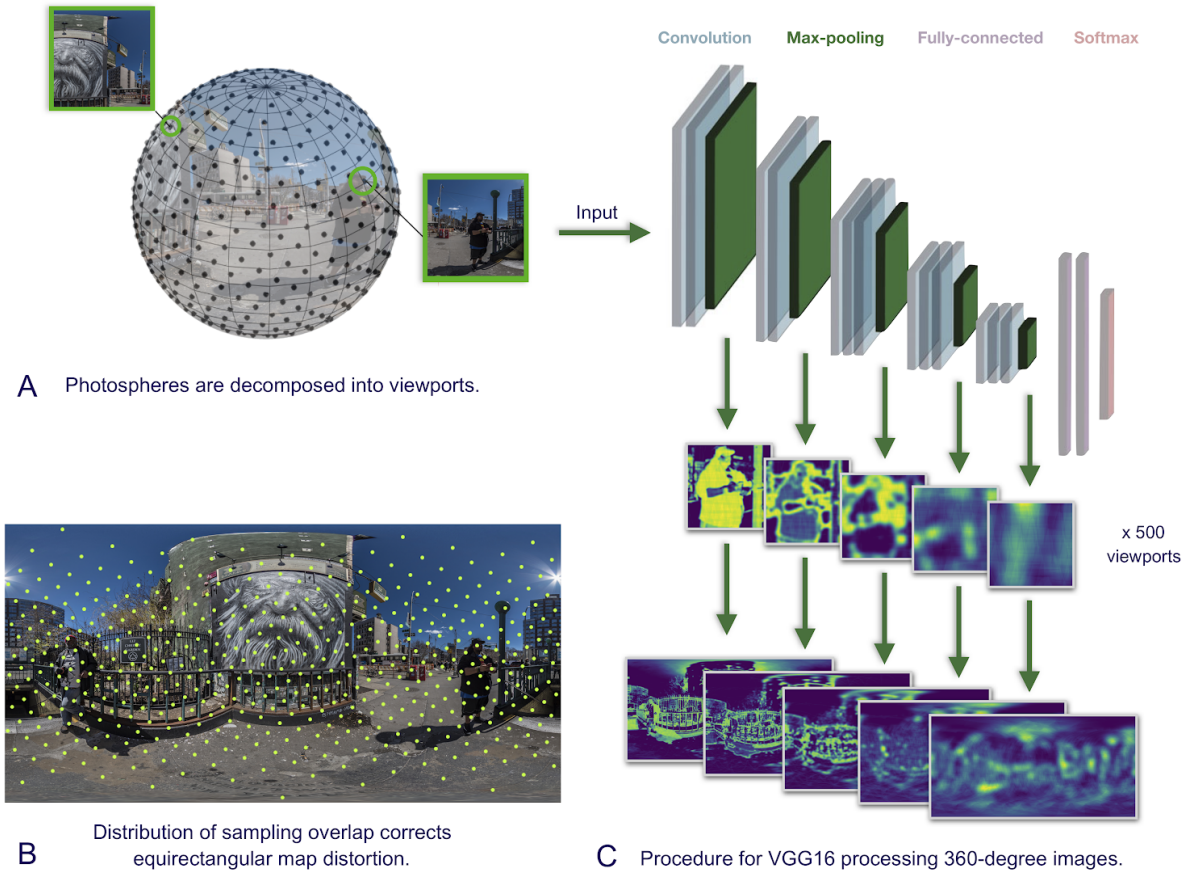


Figure 2. Photosphere-CNN processing pipeline.
A. Each scene is presented to participants as a 360-degree photosphere projected onto a virtual-reality environment in Unity. The sphere is sampled in 500 points densely around the equator and sparsely at the poles to account for the distortion of converting the photosphere into a panoramic image in Cartesian space (adapted from Haskins et al., 2020).
B. Sampling shown in equirectangular (Cartesian) space.
C. The viewport image centered around one sample point is fed as input into the CNN model (VGG16). At each max-pooling layer, we extract the network's activation in response to that viewport (adapted from O'Connell & Chun, 2018). This is repeated for all 500 viewports. These viewports are then projected back out into a panorama and aggregated in equirectangular space. Each pixel is then averaged over the number of times it was sampled.

Comparing behavioral and model data

Hierarchical layer analysis

To determine which layers of a CNN trained to recognize scenes best predict an individual's gaze behavior, we extracted activations from each of the 5 pooling layers of VGG16-Places365 (Places-CNN) in response to each experimental scene using our panoramic image feature map extraction pipeline. This yielded 5 feature maps (CNN layer maps) for each of our 60 photospheres. CNN layer maps were normalized to have zero-mean and unit variance then scaled to have values between zero and one. Finally, CNN layer maps were blurred using a multiplicative center-bias operation, which devalues activity around the poles of the scene to account for the equator bias participants consistently demonstrate in scenes, where their fixations are concentrated around the middle of the scene and rarely at the borders (Judd et al., 2009; Sitzmann et al., 2017; Henderson & Hayes, 2018).

After these normalization steps, we measured how well CNN layer maps predicted gaze behavior using Pearson's correlation coefficients. Pearson's correlation coefficients are a common metric for evaluating the predictive value of saliency models as they penalize models for both false positive and false negative predictions symmetrically (Bylinski et al., 2017). For each participant and experimental trial, gaze maps were normalized to zero-mean and unit variance then scaled to values between zero and one. Gaze maps and the corresponding scene's CNN layer maps were sampled at 1000 coordinates to control for equirectangular distortion (see Figure 2B), resulting in 1000 activation values from the gaze maps and 1000 activation values from each of the five CNN layer maps. Gaze activation values were then correlated with the CNN layer map activation values, resulting in five correlation scores per experimental trial.

For each participant, we averaged over each experimental trial's layer correlation values to obtain an average correlation score for each participant with each pooling layer of VGG16-Places365, resulting in five scores per participant. This score reflects how predictive a given layer is of the individual's gaze across all trials. We repeated this process for each participant and averaged over all participants within each diagnostic group, resulting in two sets of five correlation values: one set for how predictive each layer is of gaze behavior for individuals with ASC and one set typically developing individuals.

Local versus global attention analysis

To determine whether individuals use more local or global attention when exploring a novel scene, we compared gaze maps with CNNs trained to perform object recognition (VGG16-ImageNet; Object CNN) versus scene recognition (VGG16-Places365; Scene CNN). We compared individual gaze maps from each experimental trial to the top-pooling

layer activation generated by the Object-CNN and the Places-CNN in response to that trial's scene. To do this, we extracted just the final max-pooling layer activation, resulting in one CNN map per model (Object-CNN Map and Places-CNN Map) for each of our photospheres. Gaze and CNN maps were normalized and sampled identically as in the hierarchical layer analysis.

After normalization and sampling, gaze and CNN pool-5 maps were compared using Pearson's correlation coefficients. This was repeated for each participant and each experimental trial then averaged across all trials within the participant. This resulted in two average scores per participant: one for a participant's average correlation score with Places-CNN and one for Object-CNN. We averaged these scores across participants within diagnostic group to result in four final scores: the average correlation of gaze behavior in individuals with ASC and without ASC with the top-level activation of the Places-CNN and Object-CNN.

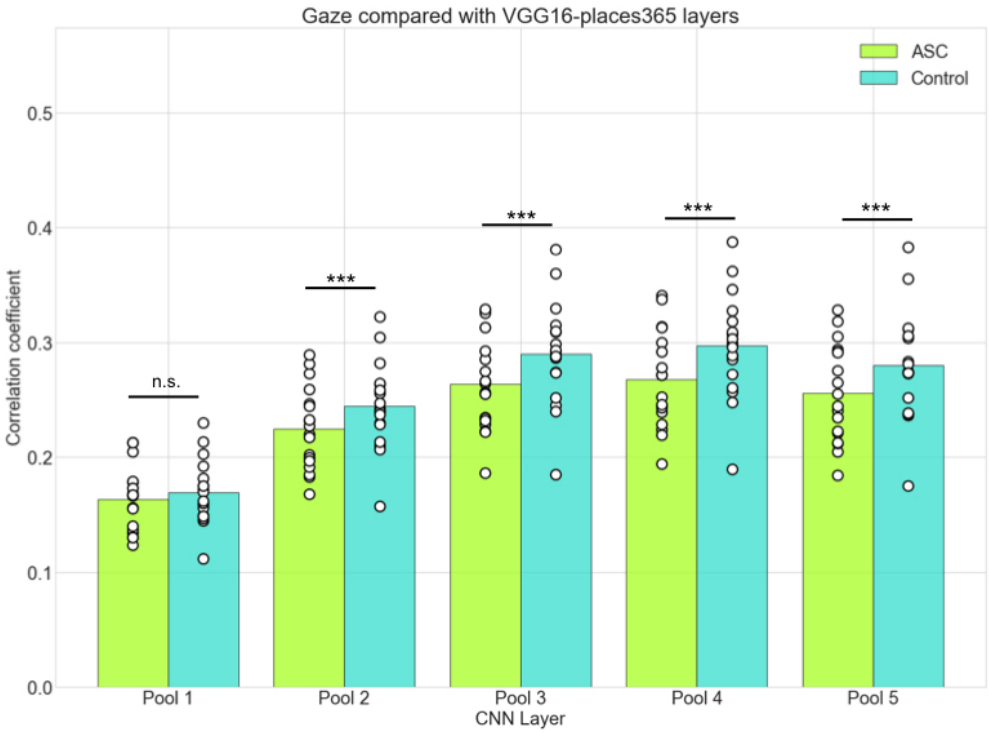
Results

How well do the layers of a scene-trained CNN predict gaze behavior?

The first analysis addressed whether different layers of a scene-processing CNN accounted for the differences in gaze behavior between individuals with and without autism. Behavioral and neuroimaging studies suggest that atypical early cortex activity could drive the visual perceptual differences hallmark of ASC. In recent years, CNNs have emerged as a model of the primate visual hierarchy. We hypothesized that earlier CNN layers would be more predictive of gaze behavior in autism, reflecting more attention to low-level features like pixel saliency or contrast, whereas later layers (which represent more semantic information) would be more predictive of typical gaze behavior.

We used Pearson's correlation coefficients to investigate how predictive each layer of the model was of gaze behavior. This resulted in five values for each trial, for each participant, per model. VGG16-Places365 differentiated between ASC and controls significantly at pool-2, -3, -4, and -5. At each level of the VGG16-Places365, controls were consistently better predicted than individuals with autism. We tested the main effects of autism diagnosis (group) and layer number with a two-factor ANOVA. For both groups, we found a main effect of layer, with correlations always increasing at higher layers (F-value = 766.95; $p < 0.0001$). We also found a main effect of group, as individuals with autism are less predicted at all layers (F-value = 430.61; $p < 0.0001$). Finally, we found a group by layer interaction (F-value = 7.0647; $p < 0.0001$; see Figure 3). To confirm these results were not driven by a few outlier participants or scenes, we modeled individual-by-layer-by-trial interactions as

random effects in a linear mixed-effects model. We included the same fixed effects of diagnosis but added random effects of individual participants and trials. We found consistent patterns in our results: main effects of group and layer with a group by layer interaction.



Analysis of Variance Table

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	7.87	7.8669	430.6063	< 2.2e-16 ***
layer	4	56.05	14.0116	766.9491	< 2.2e-16 ***
group:layer	4	0.52	0.1291	7.0647	1.11e-05 ***
Residuals	41750	762.74	0.0183		

Figure 3: CNN layer analysis results.

Top: We averaged each individual's correlation with feature maps from each layer of the CNN to obtain an average correction score per individual, per layer. These scores are represented by **open circles.** We averaged across all individuals' average correlation scores within a group at each layer to obtain one group average correlation score per layer. These are represented by the colored **bars**. Significance indicated as:

*** : p < 0.05, ** : p < 0.01, *** : p < 0.001**

A. Average individual & group gaze correlations with feature maps from VGG16-Places365.

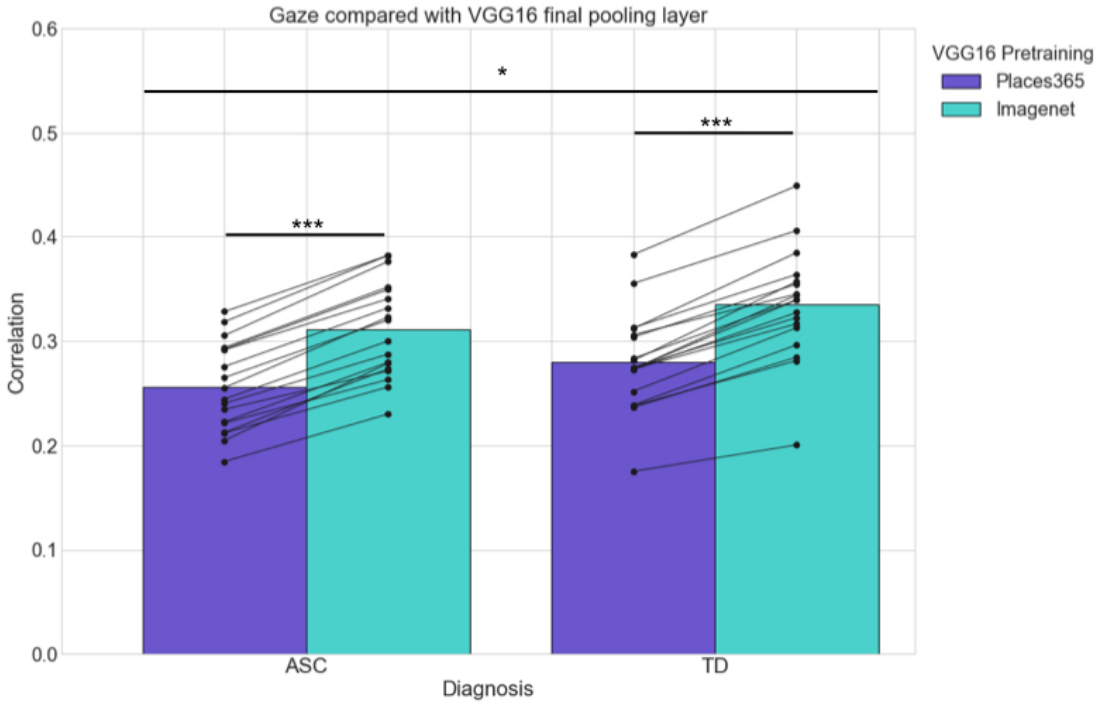
B. ANOVA table testing main effects of group and pooling layer, as well as their interactions.

Is gaze behavior in a novel scene more like an object-recognition task or a scene recognition task?

To test whether individuals exploring a novel scene allocate attention more toward local features or global features within the scene, we compared gaze maps with CNN maps of the top pooling layers across the Places-CNN and Object-CNN. These models have identical architectures but hidden weights optimized for scene classification and object classification respectively. By comparing a human's gaze behavior with the final pooling layer activity of CNNs completing object recognition tasks and scene recognition tasks within the same photosphere, we can identify which task is more comparable to the participant's active exploration of a scene.

We hypothesized that gaze behavior of individuals with ASC would be more similar to a high-level feature map of a network completing an object recognition task (VGG16-ImageNet) than one completing a scene recognition task (VGG16-Places365). We hypothesized that control behavior would reflect the reverse: their gaze behavior would be driven more by globally-informative features than locally-focused ones.

On average, individual gaze maps from ASC participants were significantly more correlated with the final VGG16-ImageNet map than with the final VGG16-Places365 map (t-statistic = 23.27; $p < 0.0001$). However, the final VGG16-ImageNet map was also more predictive of control gaze behavior than the final VGG16-Places365 map (t-statistic = 17.91; $p < 0.0001$). We modeled the effect of CNN pretraining and autism diagnosis with a two-factor ANOVA. This revealed a significant main effect of autism diagnosis ($F = 30.43$; $p < 0.0001$) as well as the CNN pretraining ($F = 149.69$; $p < 0.0001$).We also found a significant interaction of group and network, with both networks predicting control behavior better than ASC ($F = 4.86$; $p < 0.05$) (Figure 4B).



Analysis of Variance Table

Factor	Df	Sum Sq	Mean Sq	F value	Pr(>F)
group	1	0.636	0.63617	30.4283	3.674e-08 ***
network	1	3.13	3.1295	149.6851	< 2e-16 ***
group:network	1	0.51	0.5148	4.8581	0.02758 *
Residuals	4172	87.225	0.106		

Figure 4: Local/global analysis results.
Top: We averaged each individual's experimental trial correlations with pool-5 feature maps from the two models (indicated as VGG16 pretraining) to obtain an average correlation score per individual (black points). Lines connect one individual's correlation score with both models. Bars indicate the average correlation score within a group, obtained by averaging across all of the individuals' average correlation scores within diagnostic group.
Bottom: ANOVA table modeling main effects of autism diagnosis group, VGG16 network pretraining, and significant interaction of group and network.

Discussion

Hierarchical layer analysis results.
The first analysis in this study revealed that 360-gaze behavior in typically developing individuals is better predicted than in ASC at every level of the CNN. In both groups, higher levels better predict gaze behavior than lower levels, but within each level, controls are better predicted than individuals with autism. This effect is not driven by the high-level task

performed by the model (object recognition or scene recognition) -- both groups are better predicted by VGG16-ImageNet than by VGG16-Places365, and both models are more predictive of control gaze than ASC.

We had expected that later layers would diverge in their ability to predict groups rather than earlier ones, as the model reaches its prediction at its highest layer and shows the most heat around name-able objects or scene features. Instead, we found that feature maps from even the earliest layers of the CNN are more predictive of controls than of individuals with autism. This suggests that the task one performs when entering a novel space does not drive the differences in visual attention hallmark of autism. Instead, this model suggests that differences in visual perception in autism originate from changes in early visual processing and persist through all levels of processing.

Local versus global attention analysis results.

The second analysis found that gaze behavior of individuals with autism is better predicted by an object-recognition model than a scene-recognition model. We had expected this finding, as a hallmark of autistic visual perception is detail-orientation. However, we had also expected that gaze behavior of controls would be better predicted by a scene-recognition model than an object-recognition one, the reverse of our ASC prediction. We predicted this would be representative of how typically developing individuals represent a 'global percept' when exploring a novel 360-degree space. Instead, we found that control individuals were also better predicted by an object-recognition model than a scene-recognition one, and controls were better predicted by the object-recognition model than ASC individuals.

This could indicate that the visual attention of individuals with autism is less driven by semantically-salient objects in a scene than that of controls. Late layers of CNNs have the highest activation on what the model deems most significant for the task it is trying to complete. For these models, the top pooling layer is most concerned with name-able objects or scenes, which are semantically-meaningful. This suggests that ASC gaze behavior is characterized by comparatively less attention to semantically-salient features -- both social and non-social ones -- which would support the findings in Wang et al. (2015)

Conclusion

This study is the first to unite eye-tracking in virtual reality and deep learning models to study the everyday visual experience of individuals with autism. Past explorations of visual processing in ASC have been driven by theory, but this investigation is driven by data since our models are computational estimators of salience. This allows us to step away from our preconceived notions of visual processing in autism and instead peer through an unbiased

window into the sensory experience. CNN feature maps are more predictive of gaze behavior in control participants than ASC participants at every location along the processing hierarchy, regardless of the type of visual recognition task the CNN is trained to perform. These models suggest that there are differences in visual processing between individuals with and without autism, beginning early in the visual hierarchy and progressing through levels that represent high-level features.

This investigation paves the way for more computational, data-driven approaches to studying the sensory experiences of individuals with autism in real-world settings. Given the ubiquity of sensory sensitivities among individuals with autism (Robertson & Baron-Cohen, 2017), it is essential that we, as researchers, focus on developing reliable, empirical methods to understand them. By understanding and modeling the sensory differences related to autism, we can both gain insight into the neurobiology of the condition within the laboratory setting and also invest in more sensory-friendly infrastructure outside of the lab. Our approach also opens the door to understanding the perception of nonverbal individuals with autism, whose experiences we cannot understand from verbal reports, but we can gain insight into their perceptual experiences using eye-tracking data. In this report, we have only scratched the surface of what computational investigations of visual perception can teach us about autism. In the following section, we outline future extensions upon our work.

Future directions

Given the novelty of this study and the short time-frame in which this investigation was conducted, we propose a number of future avenues of investigation. We aim to address many of these in the coming months.

We would also like to understand the coherence of gaze behavior within and across diagnostic groups. Do the gaze maps of individuals with autism show less inter-subject correlation than controls do? Is one group more prone to noisy data than the other? Are there subgroups within these groups that generate more noisy data than others?

Another modification would be to our CNN map preprocessing pipeline. As shown in Figure 2C, feature maps from early pooling layers show sharp pixel contrast, whereas later feature maps are smoother. This has to do with the effective receptive field size considered in the model's layers, which increases by layer. However, this increasing smoothness gives a layer like pool-5 an inherent advantage in predicting gaze behavior. We propose smoothing our feature maps with a kernel that varies by layer in order to bring all layers into equal smoothness as pool-5. Our expectation is that this would bring all layers into equal smoothness while maintaining the specific features important to that layer.

Our final proposed future direction would be to utilize different CNN architectures for this analysis. We only tested this analysis using the VGG16 architecture. A next step would be to test this pipeline with other architectures, and if we find that another architecture better models ASC behavior, we gain insight into the visual computations defining the condition.

Acknowledgements

I would like to thank my advisors, Dr. Caroline Robertson (Dartmouth College Department of Psychological and Brain Sciences) and Dr. Leyla Isik (Johns Hopkins University Department of Cognitive Science), for their support and guidance throughout this project. I look forward to continuing this work in the future with both of you. I would also like to thank members of the Robertson Lab for their support: A.J. Haskins for her statistical prowess and R enthusiasm, Tommy Botch for co-developing the panoramic image feature map extraction pipeline, Brenda Garcia for much of the behavioral data collection, and Adam Steel for his well-timed wisdom and humor. I want to acknowledge Dartmouth's Undergraduate Research and Advising program for supporting all of my undergraduate research experiences, as well as the James O. Freedman Presidential Scholars Program and the Neukom Scholars Program through the Neukom Institute for Computational Sciences.

This study was supported by a grant from the Nancy Lurie Marks Family Foundation to Dr. Robertson and the donation of a Titan V GPU by the NVIDIA Corporation to Dr. Robertson.

References

Baranek, G. T., Watson, L. R., Boyd, B. A., Poe, M. D., David, F. J., & McGuire, L. (2013). Hyporesponsiveness to social and nonsocial sensory stimuli in children with autism, children with developmental delays, and typically developing children. *Development and Psychopathology*, *25*(2), 307–320. https://doi.org/10.1017/S0954579412001071

Baron-Cohen, S., Ashwin, E., Ashwin, C., Tavassoli, T., & Chakrabarti, B. (2009). Talent in autism: Hyper-systemizing, hyper-attention to detail and sensory hypersensitivity. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1522), 1377–1383. https://doi.org/10.1098/rstb.2008.0337

carwin. (2020). *Clcarwin/convert_torch_to_pytorch* [Python]. https://github.com/clcarwin/convert_torch_to_pytorch (Original work published 2017)

CDC. (2019, September 3). *Data and Statistics on Autism Spectrum Disorder | CDC*. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/data.html

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Collobert, R., Kavukcuoglu, K., & Farabet, C. (n.d.). *Torch7: A Matlab-like Environment for Machine Learning*. 6.

CSAILVision/places365. (2020). [Python]. MIT CSAIL Computer Vision. https://github.com/CSAILVision/places365 (Original work published 2016)

Dakin, S., & Frith, U. (2005). Vagaries of Visual Perception in Autism. *Neuron*, *48*(3), 497–507. https://doi.org/10.1016/j.neuron.2005.10.018

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (n.d.). *ImageNet: A Large-Scale Hierarchical Image Database*. 8.

Elsabbagh, M., Divan, G., Koh, Y.-J., Kim, Y. S., Kauchali, S., Marcín, C., Montiel-Nava, C., Patel, V., Paula, C. S., Wang, C., Yasamy, M. T., & Fombonne, E. (2012). Global Prevalence of Autism and Other Pervasive Developmental Disorders. *Autism Research*, *5*(3), 160–179. https://doi.org/10.1002/aur.239

Estes, A., Zwaigenbaum, L., Gu, H., St. John, T., Paterson, S., Elison, J. T., Hazlett, H., Botteron, K., Dager, S. R., Schultz, R. T., Kostopoulos, P., Evans, A., Dawson, G., Eliason, J., Alvarez, S., & Piven, J. (2015). Behavioral, cognitive, and adaptive development in infants with autism spectrum disorder in the first 2 years of life. *Journal of Neurodevelopmental Disorders*, *7*(1). https://doi.org/10.1186/s11689-015-9117-6

Fu-En.Wang. (2020). *Fuenwang/Equirec2Perspec* [Python]. https://github.com/fuenwang/Equirec2Perspec (Original work published 2017)

Gliga, T., Bedford, R., Charman, T., Johnson, M. H., Baron-Cohen, S., Bolton, P., Cheung, C., Davies, K., Liew, M., Fernandes, J., Gammer, I., Maris, H., Salomone, E., Pasco, G., Pickles, A., Ribeiro, H., & Tucker, L. (2015). Enhanced Visual Search in Infancy Predicts Emerging Autism Symptoms. *Current Biology*, *25*(13), 1727–1730. https://doi.org/10.1016/j.cub.2015.05.011

Groen, I. I., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *ELife*, *7*. https://doi.org/10.7554/eLife.32962

Güçlü, U., & van Gerven, M. A. J. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Brain's Ventral Visual Pathway. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Haskins, A.J., Mentch, J.S., Botch, T.L., & Robertson, C.E. (2020). Active Vision in

Immersive, 360° Real-World Environments. *In preparation.*

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10–10. https://doi.org/10.1167/18.6.10

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106-154.2. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1359523/

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science (New York, N.Y.)*, *310*(5749), 863–866. https://doi.org/10.1126/science.1117593

Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation Schedule, Module 4: Revised Algorithm and Standardized Severity Scores. *Journal of Autism and Developmental Disorders*, *44*(8), 1996–2012. https://doi.org/10.1007/s10803-014-2080-3

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, *114*(43), E9145–E9152. https://doi.org/10.1073/pnas.1714471114

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*, 2106–2113. https://doi.org/10.1109/ICCV.2009.5459462

Kaldy, Z., Kraper, C., Carter, A. S., & Blaser, E. (2011). Toddlers with Autism Spectrum Disorder are more successful at visual search than typically developing toddlers. *Developmental Science*, *14*(5), 980–988. https://doi.org/10.1111/j.1467-7687.2011.01053.x

Kern, J. K., Trivedi, M. H., Garver, C. R., Grannemann, B. D., Andrews, A. A., Savla, J. S., Johnson, D. G., Mehta, J. A., & Schroeder, J. L. (2006). The pattern of sensory processing abnormalities in autism. *Autism*, *10*(5), 480–494. https://doi.org/10.1177/1362361306066564

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(1), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 25* (pp. 1097–1105). Curran Associates, Inc. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). DeepGaze II: Reading fixations from deep features trained on object recognition. *ArXiv:1610.01563 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1610.01563

MATLAB 2019b (2019b). (n.d.). [Computer software]. The MathWorks, Inc.

Mottron, L., & Belleville, S. (1993). A Study of Perceptual Analysis in a High-Level Autistic Subject with Exceptional Graphic Abilities. *Brain and Cognition*, *23*, 279–309.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353–383. https://doi.org/10.1016/0010-0285(77)90012-3

O'Connell, T. P., & Chun, M. M. (2018). Predicting eye movement patterns from fMRI responses to natural scenes. *Nature Communications*, *9*(1), 5159. https://doi.org/10.1038/s41467-018-07471-9

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8026–8037). Curran Associates, Inc. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Peterson, M. F., Lin, J., Zaun, I., & Kanwisher, N. (2016). Individual differences in face-looking behavior generalize from the lab to the world. *Journal of Vision*, *16*(7), 12. https://doi.org/10.1167/16.7.12

Plaisted, K., O'Riordan, M., & Baron-Cohen, S. (1998). Enhanced Visual Search for a Conjunctive Target in Autism: A Research Note. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(5), 777–783. https://doi.org/10.1017/S0021963098002613

Robertson, C. E., & Baron-Cohen, S. (2017). Sensory perception in autism. *Nature Reviews Neuroscience*, *18*(11), 671–684. https://doi.org/10.1038/nrn.2017.112

Robertson, C. E., Thomas, C., Kravitz, D. J., Wallace, G. L., Baron-Cohen, S., Martin, A., & Baker, C. I. (2014). Global motion perception deficits in autism are reflected as early as primary visual cortex. *Brain*, *137*(9), 2588–2599. https://doi.org/10.1093/brain/awu189

Rogers, S. J., Hepburn, S., & Wehner, E. (2003). Parent Reports of Sensory Symptoms in Toddlers with Autism and Those with Other Developmental Disorders. *Journal of Autism and Developmental Disorders*, *33*(6), 631–642. https://doi.org/10.1023/B:JADD.0000006000.38991.a7

Sak, H., Senior, A., & Beaufays, F. (n.d.). *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*. 5.

Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. (2009). Vision in autism spectrum disorders. *Vision Research*, *49*(22), 2705–2739. https://doi.org/10.1016/j.visres.2009.08.005

Simonyan, K., & Zisserman, A. (2015). *VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION*. 14.

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., & Wetzstein, G. (2017). How do people explore virtual environments? *ArXiv:1612.04335 [Cs]*. http://arxiv.org/abs/1612.04335

Spencer, J. V., & O'Brien, J. M. D. (2006). Visual Form-Processing Deficits in Autism. *Perception*, *35*(8), 1047–1055. https://doi.org/10.1068/p5328

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., & Zhao, Q. (2015). Atypical Visual Saliency in Autism Spectrum Disorder Quantified through Model-Based Eye Tracking. *Neuron*, *88*(3), 604–616. https://doi.org/10.1016/j.neuron.2015.09.042

Wass, S. V., Smith, T. J., & Johnson, M. H. (2013). Parsing eye-tracking data of variable quality to provide accurate fixation duration estimates in infants and adults. *Behavior Research Methods*, *45*(1), 229–250. https://doi.org/10.3758/s13428-012-0245-6

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

YouTube. (n.d.). Retrieved February 25, 2020, from https://www.youtube.com/

Zagoruyko, S. (2020). *Szagoruyko/loadcaffe* [Protocol Buffer]. https://github.com/szagoruyko/loadcaffe (Original work published 2014)