Learning Visual Attention to Identify People with Autism Spectrum Disorder

Ming Jiang, Qi Zhao
University of Minnesota
mjiang@umn.edu, qzhao@cs.umn.edu

Abstract

This paper presents a novel method for quantitative and objective diagnoses of Autism Spectrum Disorder (ASD) using eye tracking and deep neural networks. ASD is prevalent, with 1.5% of people in the US. The lack of clinical resources for early diagnoses has been a long-lasting issue. This work differentiates itself with three unique features: first, the proposed approach is data-driven and free of assumptions, important for new discoveries in understanding ASD as well as other neurodevelopmental disorders. Second, we concentrate our analyses on the differences in eye movement patterns between healthy people and those with ASD. An image selection method based on Fisher scores allows feature learning with the most discriminative contents, leading to efficient and accurate diagnoses. Third, we leverage the recent advances in deep neural networks for both prediction and visualization. Experimental results show the superior performance of our method in terms of multiple evaluation metrics used in diagnostic tests.

1. Introduction

The 'gold standard' diagnostic process of neurodevelopmental disorders is expensive, subjective, and time-consuming. Relying on a multidisciplinary team of neurologists, pediatricians, pathologists, and therapists, it usually takes years of successive interviews and behavioral evaluations that are susceptible to interpretive biases. For example, Autism Spectrum Disorder (ASD) affects one in 68 people in the US [9]. Because of the prevalence of ASD and the limited clinical resources, early diagnoses and interventions are not widely accessible or applicable. High-functioning people on the spectrum are usually overlooked in their childhood. Therefore, quantitative and objective diagnostic tools have been in great need for decades, yet little progress has been made.

Recently, we have witnessed the fast growth of two relevant and important domains that could potentially revolutionize the diagnosis of neurodevelopmental disorders – visual attention and artificial neural networks (DNNs). The attention network is so pervasive in the brain that many neurodevelopmental disorders are associated with atypical attention towards visual stimuli [4]. However, current neurobehavioral studies of attention depend on contrived visual stimuli and structured laboratory tasks targeting particular hypotheses, which constrains the generalizability of these studies.

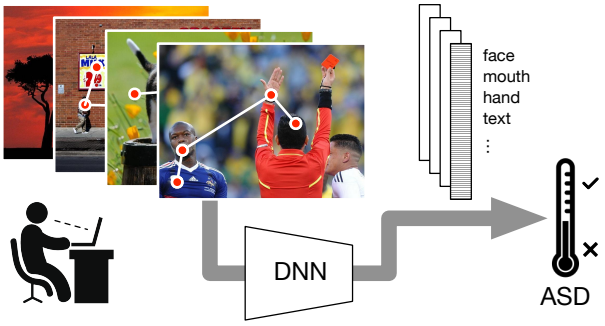

Figure 1. An overview of our experimental and classification paradigms. Subjects freely observe a selection of natural-scene images, while their eye movements being recorded. Discriminative features encoding social and nonsocial information are extracted from a deep neural network of visual attention, and integrated across images to predict the subjects' ASD risk.

To address the long-lasting issues in the diagnosis of ASD, we propose to leverage the pervasiveness of visual attention in the brain and the learning ability of DNNs. Driven by DNNs, computational models of attention have shown notable progress in recent years, encoding rich semantics and predicting human attention remarkably well in complex natural scenes. As ASD exhibits atypical gaze patterns toward social stimuli [32, 13, 31], the capability of DNNs in encoding high-level semantics and social contents is particularly suitable for identifying people with ASD.

Specifically, we propose a new diagnostic paradigm with natural scene viewing. Subjects' eye movements are recorded and analyzed quantitatively with a DNN model. With the goal of effective and efficient diagnoses, our method selects the most discriminative images by ranking their Fisher scores and learns the most discriminative features by creating and analyzing a difference map of eye fix-

ations. By using natural-scene stimuli and DNNs, the diagnostic process is completely data-driven and assumption-free.

In summary, this paper carries three major contributions:

First, because of the generality of the free-viewing task, the stimuli, and the data-driven model, our method is easily applicable to most people on the spectrum and generalizable to other neurodevelopmental disorders.

Second, selecting discriminative images and learning discriminative features boost the accuracy of ASD diagnosis while reducing the amount of data and time needed for the diagnosis.

Finally, our DNN-based model demonstrates superior performances in classifying clinical populations. The DNN visualization serves to enrich our understanding of the atypical attention in ASD.

2. Related Works

Machine learning methods have been applied to autism studies with various modalities, such as quotient-based ASD diagnostic tools [2], Magnetic Resonance Imaging data [11], acoustic data of early language [25], and kinematic data [5]. However, most of these studies rely on handcrafted features and simplified linear models that often fail to handle the complexity of human behaviors in less controlled settings. Here we briefly review the most related works using machine learning to analyze eye-tracking data for autism research and diagnosis.

Eye movements encode rich information about the attention and oculomotor control of an individual, which could help characterize ASD traits. Compared with standard diagnostic quotients, automatic and quantitative processing of eye movements will potentially lead to more objective and accessible diagnoses of ASD. Recently, Liu *et al*. [22] analyzed the gaze patterns of children with ASD in a face recognition task. They proposed a machine learning method to classify 29 children with ASD and two groups of matched controls. Despite their prominent accuracy, the face stimuli and the structured recognition task are highly dependent on the existing knowledge about ASD, limiting their generalizability to other clinical populations or young children who may fail to understand or comply with the task instruction.

To classify clinical populations with a variety of neurodevelopmental disorders (*i.e*., Attention Deficit Hyperactivity Disorder, Fetal Alcohol Spectrum Disorder, and Parkinson's disease, note that this work did not study ASD though), Tseng *et al*. [35] analyzed gaze patterns in watching short video clips. They combined gaze patterns with features from a computational model of visual attention and showed the advantages of incorporating attention-related features in identifying specific disorders. Their attention model only focused on early visual saliency, but it did not take into account high-level semantics and social information that could influence visual attention more significantly.

Our work is primarily inspired by the comprehensive study of Wang *et al*. [38]. They quantified the visual attention of high-functioning adults with ASD using a linear Support Vector Machine (SVM) and three levels of features. Feature weights were learned with recorded eye fixations and used to quantify the different attentional preferences between people with ASD and healthy controls. The main limitation of this method is the handcrafted features and the requirement of manual annotation on objects of interest. Besides, Wang *et al*. [38] only studied the differences between groups, without classification of individual subjects. Instead, to tackle these problems, our method automatically extracts image features from natural scenes using deep neural networks (DNNs) and use these features to distinguish people with ASD from healthy controls.

Deep neural networks are state-of-the-art architectures in machine learning and have been extensively used in various computer vision tasks. With sufficient training data, DNN models have overwhelmingly outperformed handcrafted features in visual recognition, detection, segmentation, and so on. Deep visual attention networks have also demonstrated effectiveness in predicting where humans look [37, 18, 21, 14, 27, 20]. Some state-of-the-art models transfer the knowledge learned from object recognition tasks to the problem of attention prediction [18] or use saliency evaluation metrics as objectives [14, 20]. Different from the related works that focused on predicting the gaze distribution of one population, this study for the first time develops a DNN model to predict the difference between two populations. Replacing the conventional fixation ground truth with a probability map that highlights between-group differences allows the DNN to learn discriminative features that are salient to one group of people but not the other.

3. Method

In this work, we utilize the rich DNN features learned from a selection of discriminative images to classify the group membership of an individual. Figure 2 demonstrates an overview of our DNN-based method. The key components of our method include eye-tracking data collection, selection of discriminative images, DNN-based feature learning, and SVM-based classification.

3.1. Eye Tracking

Eye-tracking data were collected by Wang *et al*. [38] from 20 high-functioning adults with ASD and 19 neurologically and psychiatrically healthy controls, in a free-viewing task. Subjects had normal or corrected-to-normal visual acuity. The two groups of subjects were matched on gender, age, race, and education. The ASD group all met the

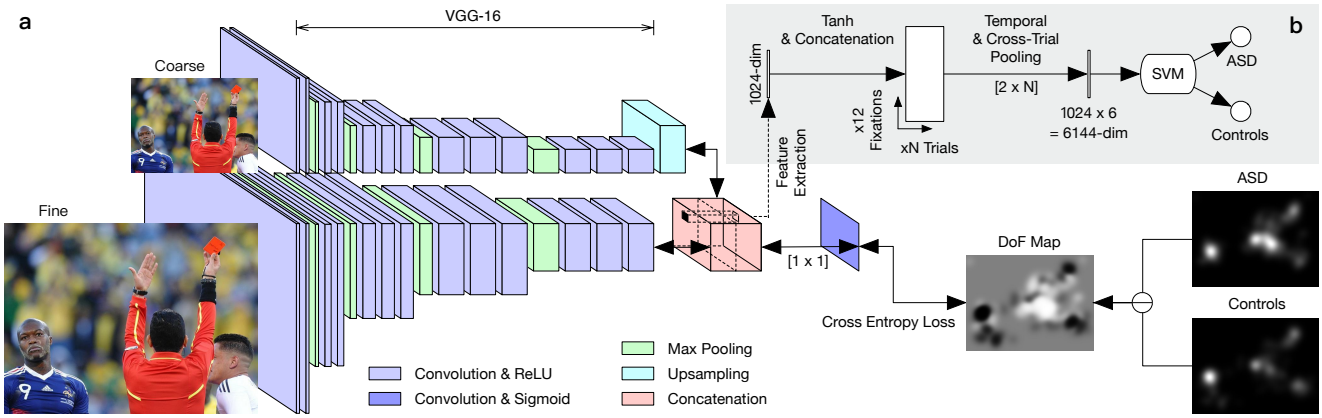

Figure 2. Overview of the proposed feature learning and classification approach. (a) Discriminative image features are learned end-to-end to predict the difference of fixation maps. (b) Features at fixated pixels are extracted and integrated across N trials to classify each eye's behavior with an SVM.

cutoff scores on the Autism Diagnostic Observation Schedule (ADOS) [23] and the Autism Diagnostic Interview-Revised (ADI-R) [24] or Social Communication Questionnaire (SCQ) when an informant was available.

Subjects observed 700 images from the OSIE [39] database. With a large variety of objects in natural scenes, these images had demonstrated advantages in understanding different aspects of visual attention, from low-level feature contrasts to high-level semantics. In the eye-tracking experiment, the images were shuffled and presented in consecutive experimental trials. Each trial began with a drift correction when subjects fixate a dot at the center of a blank screen and press a button to continue. Following the drift correction, a target image was presented for three seconds. Subjects were instructed to look at the image freely. Binocular eye movements were recorded at 300 Hz during the image viewing, with a non-invasive infra-red remote Tobii X300 eye tracker. The experiment was split into seven recording sessions of 100 trials each, with short breaks and re-calibrations between sessions. Subjects finished the experiment within two hours.

We processed the raw eye-tracking data for further analysis. Fixations and saccades were detected from the raw data with Cluster Fix [17] – a nonparametric fixation detection algorithm. We considered each saccade-fixation pair as a temporal *fixation stage* of the visual scanning process. For all trials, we discarded the data after the twelfth fixation stage.

Note that eye-tracking precision can be affected by calibration errors, environmental conditions, binocular disparity and subjects' performance. For example, in [38], the average eye-tracking error was approximately 40 pixels for ASD and 18 pixels for controls. Therefore, Wang *et al*. [38] discarded the data of the less accurate eye. However, in this work, we analyzed the data of both eyes, which prevented the model overfitting to the dominant eye and improved its robustness to eye-tracking errors. Further, in our experiments, we classified two eyes of a subject as two separate data samples. Thus, each eye was classified independently, which would allow future subjects being tested even if only monocular data were available.

3.2. Image Selection

The OSIE [39] database was originally designed for a general study of visual attention but not specific to autism. Since the number of experimental trials could directly determine the cost of diagnostic tests, we selected a subset of images from the original database that could best differentiate the gaze patterns between people with ASD and controls. We performed a feature selection based on the Fisher score method [10] to find the most discriminative images.

Selecting a discriminative subset from the 700 natural-scene images is an important component of our method. A discriminative image should maximally reveal the statistical difference between subjects from different groups while being minimally sensitive to intra-group variations. The fundamental idea of the Fisher score is to find a subset of features that are most important for the classification of subject groups. The features are selected using the principle that data points lying in the same class should be close to each other and those lying in different classes should be far from each other. Specifically, consider a feature vector $\mathbf{x}$ representing the gaze pattern of an individual, with its j-th dimension denoted as $\mathbf{x}_j$. Let $(\mu_{j+}, \sigma^2_{j+})$ and $(\mu_{j-}, \sigma^2_{j-})$ denote the means and variances of $\mathbf{x}_j$ across the ASD and the control groups, respectively. The Fisher score of $\mathbf{x}_j$ is computed as

$$F(\mathbf{x}_j) = \frac{(\mu_{j+} - \mu_{j-})^2}{\sigma^2_{j+} + \sigma^2_{j-}}. \quad (1)$$

To perform the image selection, a set of gaze features were defined, describing subjects' viewing behavior. At each fixation stage, we computed three spatial features of the fixation points (*i.e.*, their horizontal and vertical positions, and distance to the screen center) and four oculomotor features (*i.e.*, fixation duration, saccade amplitude, saccade duration, and saccade velocity). These feature vectors, concatenated over all fixation stages and images, represented an individual's viewing behavior in the eye-tracking experiment. Fisher scores were then computed at all feature dimensions and averaged per image. We selected a subset of images with the highest scores and performed feature learning and classification on the selected images only.

3.3. Learning Discriminative Features from the Difference of Fixation Maps

Learning-based models to predict human visual attention are typically optimized to differentiate salient and non-salient elements in a scene. Their ground truth is a probability distribution of eye fixations, namely the fixation map. Since our objective is to differentiate two clinical populations by what they fixated, our network is optimized upon the difference of fixation (DoF) maps, highlighting the subtle differences between two fixation maps. As demonstrated in Figure 2a, higher values in the DoF maps indicate more fixations of the ASD group, while lower ones indicate more fixations of the control group.

Specifically, a fixation map is computed with an integration of all fixations of each population, by setting fixated pixels to ones and those elsewhere to zeros. The two maps are smoothed with a Gaussian kernel (bandwidth = 1°) and normalized by the sum. Let I^+ and I^- denote the final fixation maps for the ASD and control groups, respectively. The DoF map of an image can be computed as

$$D = \frac{1}{1 + e^{-I/\sigma_I}} \tag{2}$$

where $I = I^+ - I^-$ is a pixel-wise subtraction of the two maps, and σ_I represents the standard deviation of I.

As shown in Figure 2a, our network architecture follows the design of the SALICON network [14], one of the state-of-the-arts in image saliency prediction. It consists of two parallel VGG-16 networks [33] to process the input image at two different spatial resolutions. The fine-scale network takes the original image as input, while the coarse-scale input is the image downsampled by half. With the fully-connected layers removed from the VGG-16, activations at the last convolutional layers of the two networks are scaled to the same size and concatenated. In our implementation, a new convolutional layer is added, followed by a sigmoid activation to predict the DoF map. The convolutional layer has a kernel size of 1×1 without padding, indicating a weighted linear combination. The model parameters are optimized in an end-to-end back propagation with a pixel-wise cross-entropy loss:

$$L(D, \hat{D}) = -\frac{1}{n} \sum_i D_i \log(\hat{D}_i) + (1 - D_i) \log(1 - \hat{D}_i) \tag{3}$$

where the summation is over all n output pixels. The D and $\hat{D}$ are the predicted DoF map and the ground truth, respectively. The D_i and $\hat{D}_i$ are their values at the i-th pixel. The parameters of the VGG-16 layers are initialized to the pre-trained parameters on the ImageNet [8]. The network is then fine-tuned on the current database.

3.4. Classification

We extracted the learned DNN features at each fixation position and integrated them across all experimental trials. With these features, we trained a linear SVM to classify the two populations.

Specifically, as shown in Figure 2b, we extracted the DNN responses at the fixated positions from the 1024-channel concatenated activation maps, which resulted in a 1024-dimensional feature vector at each fixation. Then, a $\tanh$ function was applied to the extracted features to transform the responses within $[0, 1)$. For each eye movement scanpath, we concatenated the feature vectors of all fixations in their temporal order, from the first fixation to the last one, and appended zeros to the end if there were less than twelve fixations. This allowed modeling the dynamic change of attention across time. To reduce the dimensionality and control overfitting, we performed a local average pooling over every two consecutive fixations (*i.e.*, size=2, stride=2). Finally, the features were averaged across all trials to represent the overall attention patterns of each eye.

The SVM was trained with the extracted features, to find a linear decision boundary with a maximum margin separating the two populations. During the testing phase, the learned SVM model made a classification for the group membership of each eye's fixation data with a corresponding confidence score that allowed us to introduce a flexible threshold to determine the final classification labels.

4. Experimental Results

In this section, we report the implementation details and a comprehensive evaluation of the proposed method.

4.1. Implementation

Training and testing. The performance of the classification was assessed with a leave-one-subject-out cross-validation. Leave-one-out cross-validation has been widely used in machine learning because of its capability of returning an almost unbiased estimate of the probability of error [36]. Particularly in this study, as the binocular eye

movements of each subject would be separately classified, to test the 78 data points from 39 subjects in our database, it was sufficient to perform 39 rounds of classification, each test with two data points and train with the remaining 76. The image selection, DNN and SVM parameters were all decided by the cross-validation.

In each run of the cross-validation, a subset of 100 images was selected based on the ranking of their Fisher scores. The DNN and the SVM classifier were trained and tested on the selected 100 images. Such a small set of stimuli could significantly reduce the cost of the eye-tracking experiment in clinical applications, because only the selected images would be used for tests.

Our model was implemented in the Python language, with the Caffe and Scikit-learn libraries. During the neural network training, we fine-tuned the network with stochastic gradient descent and a batch size of one image for a total of 10,000 iterations. The base learning rate was 10^{-6} and was divided by 10 when the error plateaued. We used a weight decay of 10^{-5} and a momentum of 0.9. No dropout was used and no data augmentation was performed, following the practice in [14]. To monitor convergence and overfitting, the network was cross-validated after every 1,000 iterations. It took approximately two hours to train the network on a GTX Titan GPU. For the SVM classification, we used an L2 regularization with the penalty parameter $C = 1$.

Evaluation metrics. We quantitatively measured the performance of our proposed approach in terms of accuracy, specificity, and sensitivity. These metrics have been widely used in almost all kinds of clinical evaluations. The accuracy measures the rate of correctly classified samples in both ASD and control groups. The sensitivity (*i.e.*, true positive rate) and the specificity (*i.e.*, true negative rate) measure the rates of correctly classified samples in the ASD and in the control groups, respectively.

We also performed a Receiver Operating Characteristic (ROC) analysis by introducing a flexible classification threshold. Test samples with SVM confidence scores above the threshold were given a positive ASD diagnosis. By varying the threshold, we plotted the ROC curves of all the true positive rates *vs*. false positive rates and computed the area under the ROC curve (AUC) as a quantitative measure of the classification performance.

4.2. Classification Performance

We present quantitative results of the proposed method in Table 1. To investigate the effects of the proposed DNN features, we trained and evaluated SVM classifiers using three different sets of features, including (1) gaze features used in image selection (see Section 3.2), (2) VGG-16 features before fine-tuning, and (3) VGG-16 features after fine-tuning. To demonstrate the effect of the Fisher-score-based image selection, we also compared the performances before and after the image selection.

	No. of Images	Acc.	Sen.	Spe.	AUC
Gaze	700	0.81	0.83	0.79	0.85
	100	0.86	**0.93**	0.79	0.88
VGG-16	700	0.85	0.83	0.87	0.89
	100	0.83	0.83	0.84	0.85
VGG-16 (fine-tuned)	700	0.85	0.83	0.87	0.89
	100	**0.92**	**0.93**	**0.92**	**0.92**

Table 1. A comparison of the classification performance on the full database and the selected 100 images, using different sets of gaze and DNN features.

As reported in Table 1, fine-tuning the proposed network on the selected images significantly improved the classification accuracy. Further, image selection increased the sensitivity of the gaze features by 10%, suggesting a reduced data variance in the ASD group, making them more distinguishable from the controls. Besides, fine-tuning increased the specificity by 8%, suggesting that the DNN encoded the visual stimuli consistently fixated by the controls. To put it another way, the controls could be better distinguished by what they looked at, rather than how they shifted their gaze. Interestingly, fine-tuning was only effective on the selected images. It is very likely that only in the images with high Fisher scores, the two groups of subjects attended to different image features that most effectively separated people with ASD from healthy controls.

Since we are the first to use eye tracking in free-viewing tasks for ASD diagnoses, no related work is directly comparable. Though targeting different groups of subjects, Liu *et al*. [22] is probably the most similar to our work in terms of modality, experimental settings, and the number of subjects. Overall, as shown in Table 2, our method reached a remarkable classification performance, better than [22] and a number of quotient-based diagnostic tools [12], including ADOS, Child Autism Rating Scale (CARS), and Autism Spectrum Disorder-Diagnostic for Children (ASD-DC). Notably, our method only takes about 10 minutes of eye tracking, comparable with [22] that requires performing a face recognition task, and the others taking 10 to 45 minutes each. With the free image viewing task, our method demonstrates the feasibility and generalizability in accurately identifying individuals with ASD.

Our performance is consistent with results from the traditional quotient-based methods for autism self-assessment. We compared the confidence scores predicted by the SVM with the clinical characteristics of the subjects. Binocular predictions of the same subject were averaged. Due to the absence of the ADOS and ADI-R evaluations for the controls, we only compared the predictions with two quantitative measurements of ASD traits in adults, the Autism

Article	Tool	Acc.	Sen.	Spe.	AUC
Falkmer *et al.* [12]	CARS	0.81	0.82	0.80	—
	ADOS	0.82	0.87	0.78	—
	ASD-DC	0.84	0.88	0.81	—
Liu *et al.* [22]	ET	0.89	**0.93**	0.86	0.89
Ours	ET	**0.92**	**0.93**	**0.92**	**0.92**

Table 2. A quantitative comparison with the most related eye-tracking (ET) study and current diagnostic tools.

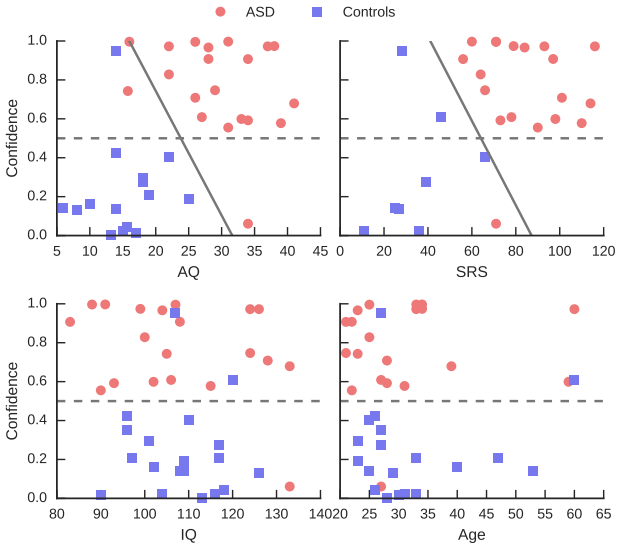

Figure 3. Classification confidence correlates strongly with the subjects' AQ and SRS scores but not IQ or age. Dashed lines indicate the classification boundary at 0.5. Solid lines indicate the optimal classification boundaries integrating our model prediction and the quotient-based evaluation.

Spectrum Quotient (AQ) [1] and the Social Responsiveness Scale-2 Adult Form Self Report (SRS). Both are widely recognized self-assessment tools for ASD in adults. As we observed, the classification confidence scores were in strong correlations with the subjects' AQ (Pearson's ρ = 0.53, p = 0.0084) and SRS (Pearson's ρ = 0.51, p = 0.0063), but had no correlation with the subjects' age (Pearson's ρ = -0.0035, p = 0.98) or their Full Scale Intelligence Quotient scores (IQ; Pearson's ρ = -0.15, p = 0.37). Our method also showed a potential to complement the quotients, improving the classification accuracy (see Figure 3).

4.3. Image Selection

Image selection has been shown effective on the classification performance (see Table 1). Below, we further elaborate its effects with detail analyses.

We first investigate the roles of the seven gaze features in image selection. Figure 4 demonstrates the change of their

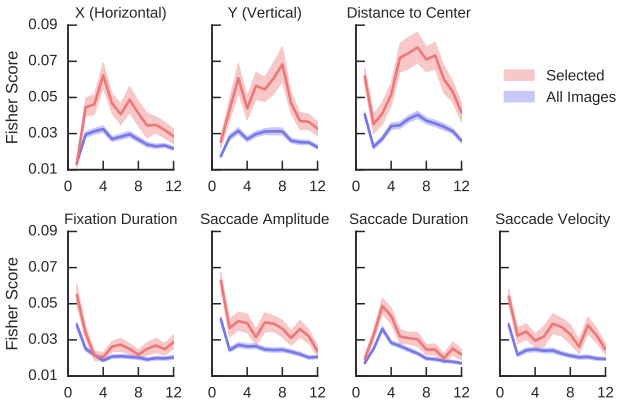

Figure 4. Fisher scores of the gaze features over time, averaged across images. Horizontal axes indicate the fixation order. Error bands indicate the standard error of the mean.

Fisher scores across fixation stages. Here the feature scores are averaged across images, but not within each of them. We can observe that after image selection, the three spatial features (*i.e.*, horizontal and vertical positions, distance to center) received more increments in their Fisher scores compared with the four oculomotor features (*i.e.*, fixation duration, saccade amplitude, saccade duration, and saccade velocity). Particularly, *distance to center* encodes the most differential information, across all fixation stages. This observation confirms previous findings that people with ASD had a stronger central preference in their eye fixations [38]. Moreover, despite the relatively low importance of the oculomotor features, they are shown to play important roles in the early fixation stages, which may suggest the impaired disengagement of attention that prevented people with ASD from saccading away from their initial fixations [7].

We compare the Fisher score with classic saliency evaluation metrics: AUC [34], shuffled AUC (SAUC) [40], linear correlation coefficients (CC) [26], information gain (IG) [19, 20], Kullback-Leibler divergence (KL) [15], normalized scanpath saliency (NSS) [29] and similarity (SIM) [16]. These metrics have been commonly used in the comparison of locations or distributions of eye fixations [3]. In general, the Fisher score agrees with all the compared saliency metrics with significant correlations. However, it has some key advantages over conventional saliency measures. Fisher scores not only measure the between-group differences but also discounts the within-group variances. Besides, in this work, oculomotor features were explicitly included, which could not be measured by the other metrics. As a result, the proposed image selection method was able to keep a high classification performance on a much smaller set of images, while the other metrics failed (see Figure 5).

Figure 6 demonstrates a qualitative comparison of images with high and low Fisher scores. Images with high

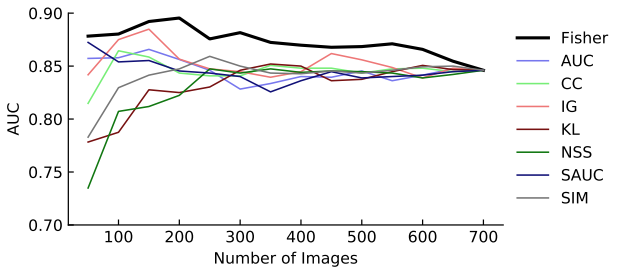

Figure 5. Comparison of classification performances (AUC) across different image selection methods. The Fisher scores are computed using the gaze features introduced in Section 3.2, while the others are location or distribution based saliency metrics.

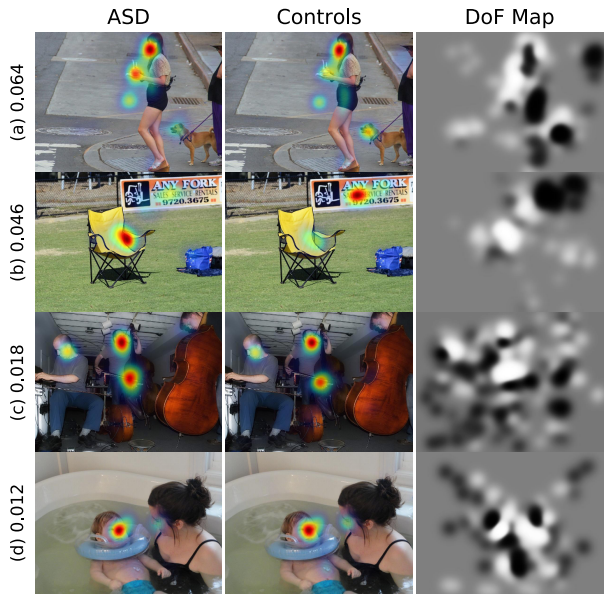

Figure 6. Example images with their Fisher scores. Fixation maps of the ASD and control groups are superimposed on the images. Their DoF maps are presented in gray scale.

Fisher scores (Figures 6a and 6b) have distinct differences in their fixation maps between the two populations. Particularly, the ASD group consistently fixated nonsocial stimuli (*i.e.*, food and chair) while controls fixated social stimuli (*i.e.*, faces and text). For images with low Fisher scores (Figures 6c and 6d), the two fixation maps are highly similar. Subtracting one with the other merely boosted eye-tracking errors and random fixations in the background. We also observed that images with higher Fisher scores tend to attract more fixations (Pearson's $\rho = 0.11$, $p = 0.005$), and the fixations are less biased towards the image center (Pearson's $\rho = 0.15$, $p < 0.001$). As suggested by Wang *et al*. [38], people with ASD tend to make fewer fixations with stronger center preference, we believe such differences are also reflected in the Fisher scores. Both qualitative and quantitative evaluations show that the proposed image selection plays an important role in our method by successfully preserving the most ASD-discriminative features.

In conclusion, our findings demonstrate that a small set of natural-scene images could reliably identify individuals with a complex and heterogeneous neurodevelopmental disorder. As a result, the diagnostic session can be conducted in a feasible length of time, and the computational model can be easily deployed in clinical or home environments.

4.4. Visualization of DNN Features

Fine-tuning a deep neural network of visual attention on the DoF maps produced highly discriminative features that significantly improved the SVM classification performance. In the upcoming analyses, we visualize the learned features and search for an interpretation of the features and their roles in existing autism studies.

We investigate the importance of the learned features by comparing their weights in the linear SVM classification. The features with the highest and lowest weights (averaged over fixations) are visualized in Figure 7. Features with positive and negative weights supported the classification between what people with ASD fixated and what the controls did, respectively. To visualize the DNN features, for each feature channel, we present 128×128 image patches at sixteen fixation points that produced the highest responses in the corresponding neuron.

As shown in Figure 7, we observed an increased lower-level saliency but decreased social attention in ASD. People with ASD were mostly distinguished by mid-level structures in a variety of nonsocial objects, *e.g.*, edges, shapes, and repetitive patterns (Figures 7a–7d). Though also looked at faces, they shifted fixations towards hair and chin (Figures 7e and 7f), suggesting a tendency to avoid eye contact [28, 6, 30]. In contrast, social features such as text, signs, human faces, upper bodies, limbs, *etc*. (Figures 7g–7l) identified controls. We also examined how the learned features corresponded to the semantic attributes defined by [39]. Notably, Figures 7g, 7i, and 7l represent neurons most activated by fixations on the three social attributes (*i.e.*, text, face, and gazed) defined in the OISE database [39]. These observations, though learned from natural scene images without any pre-assumptions, support the nonsocial preference of ASD [32, 31].

Unlike most computer vision applications, in which training and testing were performed on different images, our problem is to test new subjects. Thus the same images were used for training and testing, but with eye-tracking data of different subjects. Despite this, we performed an interesting experiment to check how the network would predict the DoF maps of novel images. Such a network would have far-reaching impact in a variety of real-world applications, such as designing autism-liked books or posters. As shown in Figure 8, the predicted difference maps well agree with

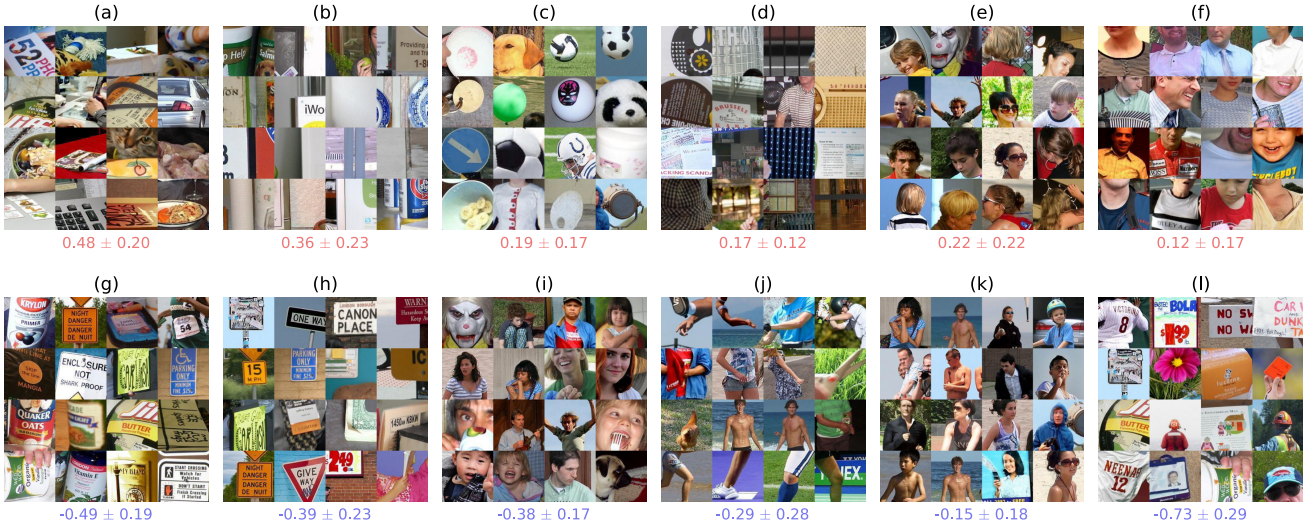

Figure 7. Features with highest and lowest SVM weights ($\times 10^{-3}$, mean±SD), visualized as arrays of image patches. Positive and negative weights represent the attentional preferences of the ASD and control groups, respectively.

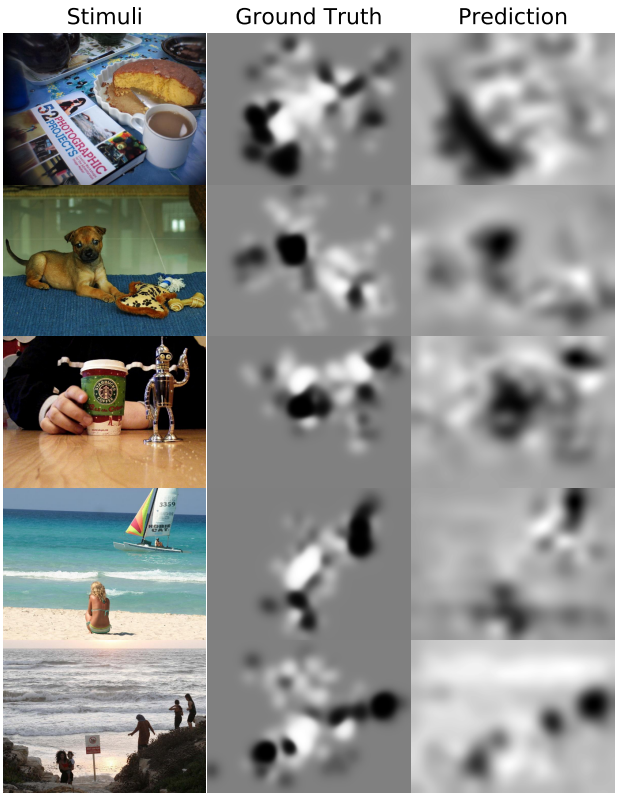

Figure 8. Predicted DoF maps on novel image examples.

the ground truth. Compared with the controls that consistently fixated social stimuli, the attention of the ASD group was allocated more randomly and difficult to predict.

5. Conclusion and Future Work

Clinical applications can significantly benefit from recent advances in machine learning, to offer more sensitive and specific diagnoses. To this end, we proposed a DNN-based model to identify people with ASD, using their eye-tracking data in free image viewing. This data-driven approach automatically learns the attention traits of ASD, without relying on any prior knowledge of the disorder. Discriminative features were learned from the differences in eye movement patterns between healthy people and those with ASD. Experimental results demonstrated promising performance with high sensitivity and selectivity. The significant predictive value of the deep neural network features and SVM classification can be valuable to support the clinical practice of diagnosing ASD. Finally, visualization of the DNN features offered insights into possible attention traits of autism with the general and ecologically relevant natural-scene images, leading further exploration towards the neuropathology of ASD.

Our proof-of-concept model has shown promising results, as a first attempt at identifying high-functioning adults with ASD. To demonstrate its generalizability, upcoming extensions of this work could include other clinical populations, such as different subgroups of ASD and other neurodevelopmental disorders. A larger database of eye-tracking data with various subject groups would also be used for developing and benchmarking future models.

References

[1] S. Baron-Cohen, S. Wheelwright, R. Skinner, J. Martin, and E. Clubley. The autism-spectrum quotient (aq): Evidence from asperger syndrome/high-functioning autism, malesand

females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1):5–17, 2001. 6

[2] D. Bone, M. S. Goodwin, M. P. Black, C. C. Lee, K. Audhkhasi, and S. Narayanan. Applying machine learning to facilitate autism diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders*, 45(5):1121–1136, 2015. 2

[3] A. Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, 2013. 6

[4] M. Corbetta and G. L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Review Neuroscience*, 3(3):201–215, 2002. 1

[5] A. Crippa, C. Salvatore, P. Perego, S. Forti, M. Nobile, M. Molteni, and I. Castiglioni. Use of machine learning to identify children with autism and their motor abnormalities. *Journal of Autism and Developmental Disorders*, 45(7):2146–2156, 2015. 2

[6] K. M. Dalton, B. M. Nacewicz, T. Johnstone, H. S. Schaefer, M. A. Gernsbacher, H. H. Goldsmith, A. L. Alexander, and R. J. Davidson. Gaze fixation and the neural circuitry of face processing in autism. *Nature Neuroscience*, 8(4):519–526, 2005. 7

[7] G. Dawson, A. N. Meltzoff, J. Osterling, J. Rinaldi, and E. Brown. Children with autism fail to orient to naturally occurring social stimuli. *Journal of Autism and Developmental Disorders*, 28(6):479–485, 1998. 6

[8] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[9] Developmental Disabilities Monitoring Network Surveillance Year and 2010 Principal Investigators and Centers for Disease Control and Prevention (CDC). Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010. *Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, DC: 2002)*, 63(2):1, 2014. 1

[10] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. 2000. 3

[11] C. Ecker, A. Marquand, J. Mourão-Miranda, P. Johnston, E. M. Daly, M. J. Brammer, S. Maltezos, C. M. Murphy, D. Robertson, S. C. Williams, and D. G. M. Murphy. Describing the brain in autism in five dimensions—magnetic resonance imaging-assisted diagnosis of autism spectrum disorder using a multiparameter classification approach. *The Journal of Neuroscience*, 30(32):10612 – 10623, 2010. 2

[12] T. Falkmer, K. Anderson, M. Falkmer, and C. Horlin. Diagnostic procedures in autism spectrum disorders: a systematic literature review. *European Child and Adolescent Psychiatry*, 22(6):329–340, 2013. 5, 6

[13] S. Fletcher-Watson, S. R. Leekam, V. Benson, M. C. Frank, and J. M. Findlay. Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia*, 47(1):248–257, 2009. 1

[14] X. Huang, C. Shen, X. Boix, and Q. Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE International Conference on Computer Vision*, pages 262–270, 2015. 2, 4, 5

[15] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009. 6

[16] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. 2012. 6

[17] S. D. König and E. A. Buffalo. A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of Neuroscience Methods*, 227:121–131, 2014. 3

[18] M. Kümmerer, L. Theis, and M. Bethge. Deep gaze i–boosting saliency prediction with feature maps trained on imagenet. *arXiv:1411.1045*, (2014):1–11, 2014. 2

[19] M. Kümmerer, T. S. Wallis, and M. Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015. 6

[20] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arxiv:1610.01563*, 2016. 2, 6

[21] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu. Predicting eye fixations using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 362–370, 2015. 2

[22] W. Liu, M. Li, and L. Yi. Identifying children with autism spectrum disorder based on their face processing abnormality: A machine learning framework. *Autism Research*, 9(8):888–898, 2016. 2, 5, 6

[23] C. Lord, S. Risi, L. Lambrecht, E. H. Cook, B. L. Leventhal, P. C. DiLavore, A. Pickles, and M. Rutter. The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3):205–223, 2000. 3

[24] C. Lord, M. Rutter, and A. Le Couteur. Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5):659–685, 1994. 3

[25] D. K. Oller, P. Niyogi, S. Gray, J. A. Richards, J. Gilkerson, D. Xu, U. Yapanel, and S. F. Warren. Automated vocal analysis of naturalistic recordings from children with autism, language delay, and typical development. *Proceedings of the National Academy of Sciences*, 107(30):13354–13359, 2010. 2

[26] N. Ouerhani, H. Hügli, R. Müri, and R. Von Wartburg. Empirical validation of the saliency-based model of visual attention. In *Electronic Letters on Computer Vision and Image Analysis*, pages 13–23, 2003. 6

[27] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016. 2

[28] K. A. Pelphrey, N. J. Sasson, J. S. Reznick, G. Paul, B. D. Goldman, and J. Piven. Visual scanning of faces in autism.

Journal of Autism and Developmental Disorders, 32(4):249–261, 2002. 7

[29] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005. 6

[30] D. M. Riby, P. J. B. Hancock, N. Jones, and M. Hanley. Spontaneous and cued gaze-following in autism and williams syndrome. *Journal of Neurodevelopmental Disorders*, 5(1):13, 2013. 7

[31] N. J. Sasson, J. T. Elison, L. M. Turner-Brown, G. S. Dichter, and J. W. Bodfish. Brief report: Circumscribed attention in young children with autism. *Journal of Autism and Developmental Disorders*, 41(2):242–247, 2011. 1, 7

[32] N. J. Sasson, L. M. Turner-Brown, T. N. Holtzclaw, K. S. Lam, and J. W. Bodfish. Children with autism demonstrate circumscribed attention during passive viewing of complex social and nonsocial picture arrays. *Autism Research*, 1(1):31–42, 2008. 1, 7

[33] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arxiv:1409.1556*, 2014. 4

[34] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, 45(5):643–659, 2005. 6

[35] P. H. Tseng, I. G. M. Cameron, G. Pari, J. N. Reynolds, D. P. Munoz, and L. Itti. High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, 260(1):275–284, 2013. 2

[36] V. N. Vapnik. An overview of statistical learning theory. 10(5):988–999, 1999. 4

[37] E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2798–2805, 2014. 2

[38] S. Wang, M. Jiang, X. M. M. Duchesne, E. A. A. Laugeson, D. P. P. Kennedy, R. Adolphs, and Q. Zhao. Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking. *Neuron*, 88(3):604–616, 2015. 2, 3, 6, 7

[39] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28, 2014. 3, 7

[40] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, 2008. 6