PREDICTING AUTISM DIAGNOSIS USING IMAGE WITH FIXATIONS AND SYNTHETIC SACCADE PATTERNS

Chongruo Wu[1*] *Sidrah Liaqat*[2*] *Sen-ching Cheung*[2] *Chen-Nee Chuah*[1] *Sally Ozonoff*[1]

[1]University of California, Davis [2]University of Kentucky

{crwu, chuah, sozonoff}@ucdavis.edu sidrah.liaqat@uky.edu sccheung@ieee.org

ABSTRACT

Signs of autism spectrum disorder (ASD) emerge in the first year of life in many children, but diagnosis is typically made much later, at an average age of 4 years in the United States. Early intervention is highly effective for young children with ASD, but is typically reserved for children with a formal diagnosis, making accurate identification as early as possible imperative. A screening tool that could identify ASD risk during infancy offers the opportunity for intervention before the full set of symptoms is present. In this paper, we propose two machine learning methods, synthetic saccade approach and image based approach, to automatically classify ASD given the scanpath data from children on free viewing of natural images. The first approach uses a generative model of synthetic saccade patterns to represent the baseline scanpath from a typical non-ASD individual and combines it with the input scanpath as well as other auxiliary data as inputs to a deep learning classifier. The second approach adopts a more holistic image based approach by feeding the input image and a sequence of fixation maps into a state-of-the-art convolutional neural network. Our experiments indicate that we can get 65.41% accuracy on the validation dataset.

***Index Terms*—** Autism Spectrum Disorders, Visual Saliency, Deep Learning

1. INTRODUCTION

Autism spectrum disorder (ASD) is defined by deficits in social and communication development and the presence of stereotyped behaviors. The natural course of ASD involves symptom onset in the first three years of life. A substantial literature demonstrates that differences between children who will later receive an ASD diagnosis and those who develop typically often emerge well before the second birthday. These differences, which include limited eye contact, response to name, shared affect, and joint attention, have been demonstrated using multiple methodologies [1, 2]. Despite this promise for early identification, the mean age of ASD diagnoses in the United States is still over age 4 [3], with less than 25% made before age 3 [4], squandering years of potential intervention when the brain is most plastic. As such, there is an urgent need in developing robust and easy-to-use ASD screening tools for infants and toddlers.

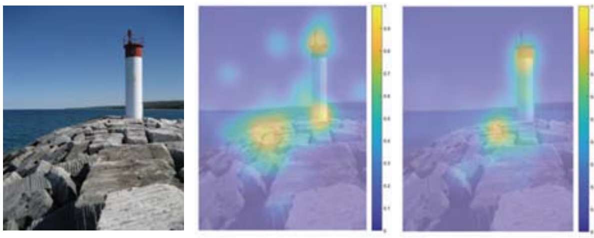
(a) Image (b) Feature map from ASD (c) Feature map from TD

Fig. 1: Sample image and corresponding saliency maps

ASD has often been associated with atypical visual attention, sometimes emerged even before the onset of the disorder [5]. Delayed disengaging attention from a previous attended location in infants has been shown to correlate with ASD diagnosis in toddlerhood [6]. Other impairments such as the inability to spread attentional resources in visual field [7] and directing attention towards less socially salient stimuli [8] have also been observed in gaze-tracking studies. These studies raise the possibility of using atypical visual attention patterns as a screening tool for ASD. While researchers are beginning to understand how these different impairments interact [9], a holistic automated diagnosis tool remain elusive.

One promising direction is to consider ASD prediction as a classification problem and use machine learning (ML) techniques to differentiate visual gaze patterns between individuals with and without ASD. As one of the grand challenges of ICME 2019, the organizers of "Saliency4ASD" have provided a dataset of images and the associated ground-truth saliency maps and gaze scan-paths of children subjects with and without ASD. A sample image along with both types of saliency maps are shown in Figure 1. One of the goals of the challenge is to propose ML models to classify ASD and typically developed (TD) viewers using gaze data.

Among the myriad of ML techniques, deep learning has

[*]The first two authors contributed equally to this paper.

978-1-5386-9214-1/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMEW.2019.00125

emerged as one of the most successful technologies in recent history. In [10], the authors studied the use of deep neural networks to identify adults with ASD using their eye-tracking data in free image viewing. Discriminative image features were learned end-to-end to predict fixation maps from which features are extracted to train a SVM for ASD classification. They have reported an impressive results of 92% accuracy on 20 high-functioning ASD and 19 typically-developed adults, which may not be directly applicable to gaze patterns from children.

In this paper, we propose two types of deep learning techniques to predict ASD diagnosis using gaze data. The first one uses a recently proposed generative model of synthetic saccade patterns called STAR-FC [11] to represent the baseline TD scanpath of a given image and combines it with the input scanpath as well as other auxiliary data as inputs to a deep learning classifier. The second approach adopts a more holistic image based approach by feeding the input image and a sequence of fixation maps into a state-of-the-art convolutional neural network. The rest of the paper is organized as follows: the details of the two proposed approaches are provided in Sections 2 and 3. Experimental results are presented in Section 4 with discussions.

2. SYNTHETIC SACCADE APPROACH

One of the commonalities reported in literature among people on the autism spectrum is that their gaze pattern when looking at different objects differs from that of neurotypical individuals. This is the main motivation behind the present approach where a deep neural network has been jointly trained on real scan path in conjunction with synthetic scan path generated by STAR-FC. Since STAR-FC is trained on general population, it is assumed that it models the scan path of TD subjects.

2.1. STAR-FC

Proposed in [11], STAR-FC is a multi-saccade generator which produces temporally ordered human-like sequences of fixation locations for a given image. The input image is first centrally fixated, followed by a retinal transform that provides anisotropic blurring centered at the current fixation point. A conspicuity map is then calculated by combining a peripheral stream dominated by low-level features and a central stream based on high-level features identified by deep networks. To identify the next fixation, a priority map is formed by combining the conspicuity map and an inhibition of return mechanism based on all previous fixations. The next fixation point is finally selected by maximing the priority map and the whole process repeats. An example of the synthetic scanpath alongside with scanpaths from both ASD and TD subjects are shown in Fig. 2.

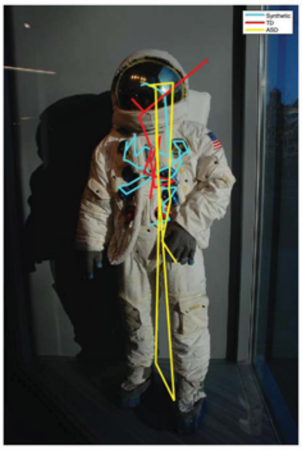

Fig. 2: TD (red), ASD (yellow) and synthetic (cyan) fixation plot on an image

2.2. Feature generation

The dataset used in the challenge [12] consists of scan paths based on location coordinates of all the fixation points and their duration. These real scan-paths are individually aligned with the corresponding synthetic scan points from STAR-FC using dynamic time warping so that the overall distance between the two paths is minimized while the order is respected. For the present model, ten points from the real as well as synthetic scan-path are used as feature. Before the alignment, the first fixation of the synthetic scan is removed as it is always at the center of the image. Since the dimensions of the example images are not the same, all the features are scaled according to the image dimensions and normalized.

Some other statistics from the real scan-path data were also used as features. These include total duration of viewing, the total number of fixation points, the mean and variance of the duration of the fixation points. The inclusion of the duration information is to reflect the possible delay effect in attention shifting among ASDs. Additionally, three different distance measures namely Dynamic Time Warping (DTW), Hausdorff distance and Frechett distance are computed between the normalized real and synthetic scan-path pair. These are common trajectory based distance measurements used in comparing scanpaths [11].

2.3. Architecture and Implementation Details

Two independent, fully connected dense networks (FCN) have been trained as separate models.

The first model has a reduced set of high level features as input namely the duration and the total number of fixation points by a single user, the mean and variance of fixation points and the three distance measures (DTW, Hausdorff and Frechett). The dimension of the input feature vector is 7. There are eight fully connected layers with seven neurons in each layer. The output layer is one-hot encoded. Batch normalization has been applied at the input layer. Selu activation has been applied on all intermediate layers. The model is trained for 200 epochs using binary cross-entropy loss with L2 regularization and Adam optimizer for with a batch size of 128. The resulting model has 492 trainable parameters.

The second model is a deeper network with 10 layers that takes as input all available information. The length of the 1D feature vector is 47 with 20 points from a real scan-path and 20 points from synthetic scan path consisting of x- and y-coordinates of fixations, in addition to the 7 high level features that are used by the previously described model. This feature vector is passed to a dense fully-connected networks having dimensions of the subsequent layers increasing to 128, 256, 512, 1024 neurons and then decreasing to 512, 256, 128, 64, 16. The output is one-hot encoded. Batch normalization and dropout of 0.3 has been applied at all the hidden layers which are activated with selu activation. The model is trained for 300 epochs on binary cross-entropy loss with Adam optimizer having learning rate of 0.001 with a batch size of 32. The resulting model has 1,402,560 trainable parameters.

3. IMAGE BASED APPROACH

ASD and TD subjects may have different behaviors when looking at the same image. In the second approach, our ASD prediction is based on the image they look at and the gaze data $D = \{(x, y, d)\}$ where (x, y) denotes the location of each data point and d denotes the time duration that the subject looks at that point.

In order to take these two sources into account, we use a network with two branches to perform this task. One branch is used to extract features of image, and the other one is to process the data points. We then fuse these two features and do prediction. Fig. 3 illustrates our model architecture. Details will be explained in the following subsections.

3.1. Data Format

Each data point consists of information on location and duration. To improve the exploitation of data points by the neural network, we convert data points to image format which is similar as the format in the human keypoint prediction task.

Each data point $p = (x, y, d)$ is represented as one image channel. The size is the same as the size of input image in the first branch. All values in this channel are zero but the value at location (x, y) is d. Instead of directly reading the coordinates of the location, our neural network could easily recognize the location. Since the max number of data points for each subject in the data set is 33, we set the number of channels of input as 33. The channel number corresponds to the order of data points. If number of data is smaller than 33, all values in the rest channel would be zeros by default.

After converting, each channel only contains one non-zero value, which is very sparse. To avoid this problem, we apply Gaussian Filter to each channel, or replicate the duration value to a radius of 5 pixles around its original location.

3.2. Architecture

In both branches, Resnet[13] is used as our backbone. For the first branch, we use a pretrained resnet18 and fix its parameters since it was already able to extract good features for natural images. Each image will be mapped to a 512-dimension vector. For the second branch, another pretrained resnet18 is used but the first convolution layer is replaced with a new one with 33 input channels. Each series of data points are also transformed to a 512-dimension vector. The concatenated 1024-dimension vector, is fed into a classifier to identify whether a subject has ASD. The total number of trainable parameters is around 11.3 million.

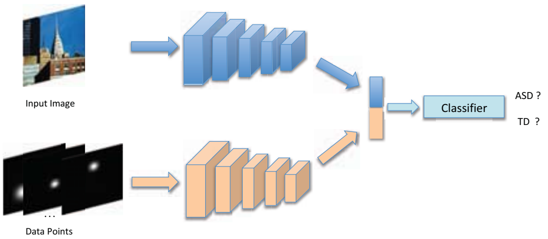

Fig. 3: An overview of our model architecture. It consists of two branches. One of them is using Resnet to extract features of images. The second one is for data point. These two features are concatenated and then fed into classifier.

3.3. Implementation Details

The data provided contains 300 images. There are 6050 samples from TD and ASD subjects. We split them into training and validation dataset (80% vs 20%). Due to the limited GPU memory, input for both branches are resized to 224 × 224.

Since dataset is small, we apply some data argumentation methods to alleviate the overfitting problem. We jitter the color for the input image. We also add random noise to the location of data points, shifting them by 20 pixels at most, and onto the duration values by multiplying them with a random coefficient from the range [0.75, 1.25]. When applying the Gaussian filter, we also use different sigma values. In each iteration, the sigma value is randomly selected from range [0.1, 2]. We also horizontally flip the images. However, we do not include any affine transformations as they may cause some data points to be out of the image range. Binary cross entropy is used as the loss function.

Method	Accuracy	Sensitivity	Specificity	AUC
Synthetic Saccade (small)	65.41%	0.66	0.65	0.69
Synthetic Saccade (full)	63%	0.69	0.66	0.66
Image -based	61.62%	0.60	0.64	0.63

Table 1: Results of our methods on validation dataset

Method	Accuracy	Sensitivity	Specificity	AUC
Synthetic Saccade (small)	54.15%	0.741	0.351	0.546
Synthetic Saccade (full)	53.88%	0.807	0.282	0.545
Image -based	55.13%	0.635	0.471	0.553

Table 2: Results of our methods on the test dataset

The model is trained by using Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is 2e-4. Batch size is set to 128. We train the whole network for 30 epochs. Our model is implemented by Pytorch and ran on Titan XP GPU.

4. EXPERIMENTS

The dataset[12] consists of scanpath data, including location and duration, from children with both ASD and TD when they look at images. 300 different images are used in the experiment. Each image is viewed by 14 ASD children and 14 TD children. Each children view one image with original full resolution for 3 seconds. The age of children with ASD lies in the range from 5 to 12 years old (8 years old on average). The scanpath data is collected by Tobii T120 eye tracker with 17-inch monitor. In order to compare the performances of the three models, we have separated all the images and the associated scanpaths into two groups: the training group has 240 images and the associated 5542 scanpaths while the testing group has 60 images with 1411 scanpaths. The models are trained on the samples from the training group and the results in Table 1 are based on testing these models on the testing samples.

During the inference, 100 images and their scanpath data are provided by the organizer, but without groundtruth label. Table. 2 shows our performance on this test dataset. The performance gap may be caused by the imbalanced distribution of training and test dataset and lacking of training images.

5. REFERENCES

[1] Grace T Baranek, "Autism during infancy: A retrospective video analysis of sensory-motor and social behaviors at 9–12 months of age," *Journal of autism and developmental disorders*, vol. 29, no. 3, pp. 213–224, 1999.

[2] Amy M Wetherby, Juliann Woods, Lori Allen, Julie Cleary, Holly Dickinson, and Catherine Lord, "Early indicators of autism spectrum disorders in the second year of life," *Journal of autism and developmental disorders*, vol. 34, no. 5, pp. 473–493, 2004.

[3] Jon Baio, "Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2010," 2014.

[4] R Christopher Sheldrick, Melissa P Maye, and Alice S Carter, "Age at first identification of autism spectrum disorder: an analysis of two us surveys," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 56, no. 4, pp. 313–320, 2017.

[5] Lori-Ann R Sacrey, Vickie L Armstrong, Susan E Bryson, and Lonnie Zwaigenbaum, "Impairments to visual disengagement in autism spectrum disorder: a review of experimental studies from infancy to adulthood," *Neuroscience & Biobehavioral Reviews*, vol. 47, pp. 559–577, 2014.

[6] Mayada Elsabbagh, Janice Fernandes, Sara Jane Webb, Geraldine Dawson, Tony Charman, Mark H Johnson, British Autism Study of Infant Siblings Team, et al., "Disengagement of visual attention in infancy is associated with emerging autism in toddlerhood," *Biological Psychiatry*, vol. 74, no. 3, pp. 189–194, 2013.

[7] Tania A Mann and Peter Walker, "Autism and a deficit in broadening the spread of visual attention," *Journal of Child Psychology and Psychiatry*, vol. 44, no. 2, pp. 274–284, 2003.

[8] Warren Jones and Ami Klin, "Attention to eyes is present but in decline in 2–6-month-old infants later diagnosed with autism," *Nature*, vol. 504, no. 7480, pp. 427, 2013.

[9] Luca Ronconi, Maria Devita, Massimo Molteni, Simone Gori, and Andrea Facoetti, "Brief report: When large becomes slow: Zooming-out visual attention is associated to orienting deficits in autism," *Journal of autism and developmental disorders*, vol. 48, no. 7, pp. 2577–2584, 2018.

[10] Ming Jiang and Qi Zhao, "Learning visual attention to identify people with autism spectrum disorder," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3267–3276.

[11] Calden Wloka, Iuliia Kotseruba, and J. K. Tsotsos, "Saccade sequence prediction: Beyond static saliency maps," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[12] X. Min Z. Che Y. Fang X. Yang J. Gutirrez P. Le Callet H. Duan, G. Zhai, "A dataset of eye movements for the children with autism spectrum disorder," in *ACM Multimedia Systems Conference (MMSys19)*, June 2019.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.