SP-ASDNET: CNN-LSTM BASED ASD CLASSIFICATION MODEL USING OBSERVER SCANPATHS

Yudong Tao, Mei-Ling Shyu

Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL, USA
{yxt128, shyu}@miami.edu

ABSTRACT

Autism spectrum disorder (ASD) is one of the common diseases that affects the language and even the behavior of the subjects. Since the large variations in the symptoms and severities of ASD, the diagnosis becomes a challenging problem. It has been witnessed that deep neural networks have been widely used and achieve good performance in various applications of visual data analysis. In this paper, we propose SP-ASDNet which utilizes both convolutional neural networks (CNNs) and long short-term memory (LSTM) networks to classify whether an observer is typical developed (TD) or has ASD, based on the scanpath of the corresponding observer's gaze at the given image. The proposed SP-ASDNet is submitted to 2019 Saliency4ASD grand challenge and achieves 74.22% accuracy for validation.

***Index Terms*—** Autism Spectrum Disorder (ASD), Deep Neural Network (DNN), Long-Short Term Memory (LSTM) Network, Saliency Prediction

1. INTRODUCTION

Autism spectrum disorder (ASD) is a developmental disease that affects the communication, behaviors, and social skills of the subjects and reported to occur for about 1 in 59 children, according to the estimation of Autism and Developmental Disabilities Monitoring Network [1]. Therefore, it becomes a critical challenge to diagnose and screen children with ASD efficiently and effectively. A precise and timely diagnosis can be helpful to mitigate the ASD effect for the children. The current procedure of a comprehensive ASD evaluation requires the assessment from well-trained specialists, which is not available in less-developed regions. Hence, it is more desirable to develop techniques that can objectively assess whether children with or without ASD. These techniques can help the early-stage detection of ASD and potentially monitoring the efficiency of the ASD remediation protocol [2].

It has been reported that children with ASD can have atypical patterns in gaze perception [3], which is caused by the disruption to early visual processing of the children with ASD. Therefore, the scanpath which characterizes the locations and durations of the gazes, responding to the given stimulus, becomes a useful cue to determine whether the observer has ASD or not.

In 2019 Saliency4ASD grand challenge, the dataset containing scanpaths of children with and without ASD for 300 images is released [4], which provides a benchmark to evaluate the saliency prediction models for ASD children and the scanpath-based ASD classification algorithms. As one of the tracks in the grand challenge, the challengers are required to classify the ASD and normal observer based on one scanpath and the corresponding image stimulus. In this paper, we propose SP-ASDNet which integrates convolutional neural networks (CNNs) and long short-term memory (LSTM) networks [5] to accomplish the ASD classification task. In addition, the pre-trained saliency prediction model, SalGAN [6], is leveraged as one component to generate the inputs of the proposed networks along with a data pre-processing procedure.

The rest of this paper is organized as follows. Section 2 introduces the related work in deep neural networks and recent advances in saliency prediction models. Then, our proposed SP-ASDNet for ASD prediction and the overall framework are presented in Section 3. Section 4 explains some details about the datasets and the model configuration and shows the experimental results of the proposed model. Section 5 concludes our findings and discusses the future work.

2. RELATED WORK

2.1. Deep Neural Networks

Deep neural networks have become one of the most effective techniques for various applications, especially when the training dataset is large [7]. As one type of deep neural networks, CNNs have become one of the most important techniques in visual data processing since AlexNet [8] was proposed in 2012 and achieved significant improvements in ImageNet competition. It has been proven to be one of the most effective techniques to learn features from images.

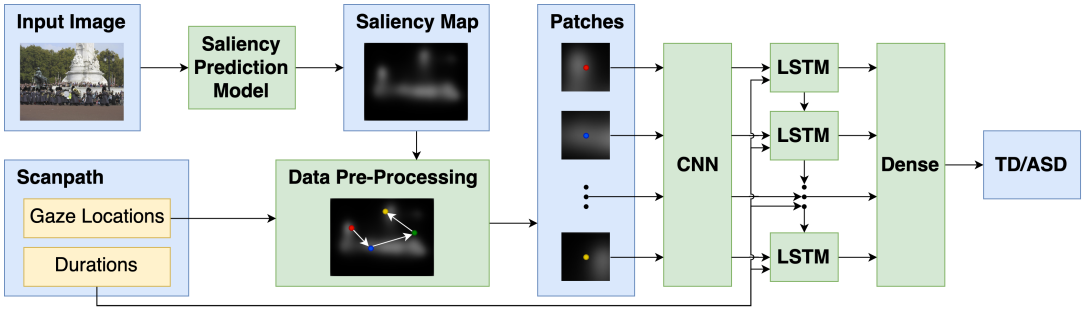

Fig. 1. The proposed SP-ASDNet framework.

On the other hand, recurrent neural networks (RNNs) have been widely applied to natural language processing (NLP), speech recognition, and other sequential data analysis, and achieved much better results than the conventional models. LSTM networks [5], as one of the most important variants of RNNs, have shown robust performance and been effectively used to all kinds of sequential data, including videos, audios [9], and scanpaths [10].

2.2. Saliency Prediction

A saliency map describes the likelihood of each pixel in the image attracting the observer. The success of deep neural networks on visual data analysis has triggered the development of deep saliency models since 2014 [11]. It has been illustrated that the deep saliency models perform much better than the conventional models.

Among all the static deep saliency models which take static images as the input, some of the recent advances will be introduced in details. One of the approaches to further enhance the prediction performance is to integrate the encoder-decoder structure into the deep saliency prediction model. EML-Net [12] implements the encoder-decoder structure in fully convolutional neural (FCN) networks which are trained separately. An ensemble of CNNs can be trained with different datasets to form the encoder to improve the overall performance. Meanwhile, SalGAN [13] leverages the architecture of the generative adversarial networks to train a generator with the encoder-decoder structure to estimate the saliency map of the images. The adversarial training process can thus boost the performance of convention encoder-decoder structure by constructing more adequate supervision signals for the model training. More recently, the attention mechanism has also been incorporated. For example, Saliency Attentive Models (SAM) that combine FCN with an attentive recurrent neural network were proposed [14].

2.3. Visual Attention for People with ASD

The hypothesis that people with ASD have different patterns of visual attention from typical developed (TD) people has been investigated and validated with the eye-tracking examination [15]. By incorporating recent advances in image recognition and saliency prediction, the researchers have been able to discover some atypical patterns from visual attention data from people with ASD [16]. These findings illustrate a prospective approach to diagnose ASD, which can potentially help the early-stage diagnosis of ASD and provide an objective measurement for ASD diagnoses. Following this track, Duan et al. has built a saliency prediction dataset for children with ASD, consisting of eye-tracking data with 13 children for 500 images [17]. As a pioneering work in utilizing visual attention data to determine whether a person has ASD or not, [18] proposed a deep neural network architecture to perform ASD classification based on a set of images. Meanwhile, an algorithm to select images with the most discriminative contents was proposed, which can be an effective tool to analyze the visual attention patterns of people with ASD.

3. SP-ASDNET

In this section, the details of the proposed SP-ASDNet framework is introduced. In general, the scanpath-based ASD classification task is to predict whether a person has ASD or not based on an image I and his/her gaze scanpath of the image $SP(I) = [(p_1, t_1), (p_2, t_2), \ldots, (p_N, t_N)]$, where p_i is a position of the image where the subject looks at, t_i is the duration of the gaze, and N is the total number of the recorded fixation locations. For this purpose, we propose to leverage the deep neural network to extract latent features of the scanpath and make the decision. Figure 1 depicts our proposed SP-ASDNet classification model which is based on the CNN-LSTM architecture using the observer scanpaths. As can be seen from Figure 1, a pre-trained saliency prediction model is first used to generate the reference saliency map of normal people for the given image. Then, a sequence of image patches of the predicted saliency map is generated based on the given scanpath, and is fed into the proposed SP-ASDNet for ASD classification to predict whether the subject has ASD or not.

3.1. Saliency Prediction

In SP-ASDNet, a pre-trained model estimating the saliency map S of the given image I is used. More specifically, the SalGAN model [13] is applied, which follows the architecture of the generative adversarial networks, i.e., composing of a generator and a discriminator. As shown in Figure 2, the generator takes the input image I to estimate the saliency map S, while the discriminator tries to differentiate the ground truth and the estimated result. The binary cross entropy is used to train the generator and the adversarial loss is used to train the discriminator. Such an adversarial training process can improve the generative model and enforce the generated results similar to the ground truth, which leads to the state-of-the-art performance of SalGAN in saliency map prediction.

In the Saliency4ASD grant challenge, we use the generator in the SalGAN model[1] pre-trained on the SALICON dataset [6] to generate the reference saliency map for the ASD classification. This component can be replaced by any other saliency prediction models, though the influence of saliency prediction performance on the final classification performance has not exhaustively tested. The SALICON dataset was collected from subjects over a large age range and thus not perfectly reflected the saliency distribution of the children. However, no public large-scale child saliency dataset is available, with which we can train the model. If sufficient training samples can be provided, the model can also be trained based on the ground truth saliency maps from children, which should provide a closer data distribution to the test sample for ASD classification and provide better classification performance.

3.2. Data Pre-Processing

In order to collect more comprehensive information from the saliency maps of the images and the scanpaths, the proposed framework extracts features from the patches of the saliency maps around the fixation locations in the scanpaths. Let $p_i = (x_i, y_i)$ be the i-th fixation location in the given scanpath, where $x_i \in [1, H]$ and $y_i \in [1, W]$ are the coordinates, and W and H are the width and height of image I, respectively. The i-th patch of the saliency map is generated by cropping the reference saliency map S, i.e., $P\left(S, p^{lt}, p^{rb}\right)$, where $p^{lt} = (x_i - a, y_i - a)$ is the coordinates of the left-top corner of the patch in S, $p^{rb} = (x_i + a, y_i + a)$ is the coordinates of the right-bottom corner, and $2a+1$ is the length of the square patch. It is assumed that the subjects should not look at any place outside the region of the image, so zero-padding is applied whenever the patch includes a region outside of the reference saliency map. The duration t_i of p_i is left aside and integrated into the feature vector of each patch before feeding into the LSTM model.

[1] https://github.com/imatge-upc/saliency-salgan-2017

Table 1. The design of the proposed CNN structures

Model	Type	#Channel	Kernel Size	Stride
A	Conv	32	7	1
	AvePool	32	2	2
	Conv	32	3	1
	Conv	16	3	1
	AvePool	16	2	2
	Dense	1024	-	-
B	Conv	32	3	1
	AvePool	32	2	2
	Conv	16	3	1
	AvePool	16	2	2
	Dense	1024	-	-

3.3. SP-ASDNet Models

After the patches are generated, a visual feature vector f_i^v is computed by a shallow CNN. The design of the CNN structures in SP-ASDNet is shown in Table 1. Two variants of the CNN structures are used and tested with the provided training dataset, namely models "A" and "B". The column "Type" shows the type of each layer, where "Conv" means the convolutional layer, "AvePool" means average pooling, and "Dense" means the fully connected layer. Each row in the table is a layer in the model and the outputs of the layers are used as the inputs of the layer below. Both models "A" and "B" produce a 1024-dimension visual feature vector f_i^v.

The feature vector of the i-th patch $f_i = [f_i^v, t_i]$ is then generated by concatenating the visual features and the duration. Thus, the feature of each patch is a vector of 1025 dimensions. Afterwards, the feature vectors of all patches of the given scanpath are fed to the two-layer LSTM networks of length L with a dense layer to classify whether the subject has ASD or not at the end. The LSTM networks are applied to capture the temporal information and sequential dependency of the scanpaths, and to integrate the feature vector of each patch together. In the case that the length of scanpath $N < L$, the zero-padding feature vectors are applied to enforce the same length of the inputs. The max-pooling layer is applied to collect the outputs of LSTM and produce a compact feature vector to the last dense layer for classification. Batch normalization layers [19] are added after each convolutional layer and the dropout layers [20] are added to each dense layer to avoid model overfitting.

4. EXPERIMENTS

4.1. Dataset

In order to train the ASD classification networks, the eye movement dataset for ASD children [4] provided by the Saliency4ASD grand challenge organizer is used. The training dataset includes 300 images, where each image is ob-

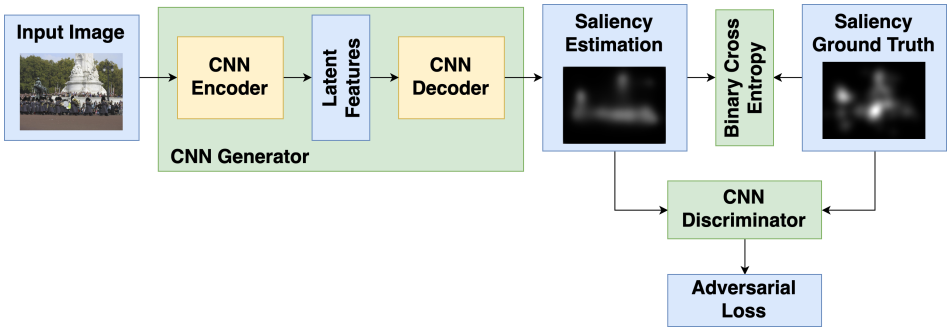

Fig. 2. SalGAN for saliency prediction [13].

Table 2. Classification accuracy on the validation dataset

Model	#Visual Features	Acc. without BN	Acc. with BN
A	512	-	58.60%
	1024	73.21%	**74.22%**
	2048	-	68.89%
B	512	-	60.79%
	1024	73.48%	74.19%
	2048	-	68.05%

served by 14 TD children and 14 ASD children. Since some of the children might not look at the locations inside the images, some of the images have less than 14 scanpaths from the TD or ASD groups.

In the Saliency4ASD grand challenge, we use the first 80% images (i.e., images 1-240) and the corresponding scanpaths to train the proposed model, while the images 241-300 and their corresponding scanpaths are used as the validation dataset to select the model for testing.

4.2. Environment Setup

For track 2 of the Saliency4ASD grand challenge, the patch size is selected as 225, i.e., $a = 112$, and the length of the LSTM is configured as 40. Both the model with and without the batch normalization layers are trained, while the ones with batch normalization shows much better performance for the validation dataset.

The training process of the proposed ASD classification model is performed on NVIDIA P100 GPU with 16GB memory. The batch size is set to 8. Adam optimizer is used to train the networks with the learning rate $lr = 1e - 5$ and the momentum of 0.9. Each model is trained for 30 epoches and the model performing the best on the validation dataset is selected for testing. During the evaluation, the weights of the model are fixed and no dropout is applied.

Table 3. Classification accuracy on the test dataset

Model	Acc.	Rec.	Pre.	F1
Model A w/t BN	0.5566	0.8771	0.5319	0.6577
Model B w/t BN	0.5739	0.5936	0.5684	0.5676
Model B w/o BN	0.5790	0.5921	0.5626	0.5697

4.3. Experimental Results

Based on the aforementioned dataset and setup, the performance of the proposed models in terms of the validation accuracy on the validation dataset can be found in Table 2, where "Model" column refers to the CNN design, "#Visual Features" column refers to the number of visual features M generated from the CNN (i.e., the number of channels of the last dense layer), the "Acc. without BN" column shows the accuracy of SP-ASDNet without batch normalization, and the "Acc. with BN" column shows the accuracy of SP-ASDNet with batch normalization. As we can see in the table, the performance with batch normalization is slightly better than those models without batch normalization, and the best 74.22% accuracy on the validation dataset is achieved.

We also investigate how the performance is affected when M varies. As can be seen from Table 2, the best performance is achieved when $M = 1024$. We think that the performance degrade when $M = 2048$ is caused by model overfitting while the visual features do not sufficiently represent the patch of the saliency map when a smaller M is applied. In the evaluation stage, three models (namely model-A with batch normalization, model-B with batch normalization, and model-B without batch normalization) are submitted to the grand challenge organizers and their performance on test dataset is shown in Table 3 where the accuracy, recall, precision, and F1 scores of the classification are listed. The model performance on the test dataset is much lower than that of the validation dataset, which could indicate the submitted models being overfitted. Therefore, the structure within the visual attention should be further investigated to improve the model and to mitigate the overfitting problem. More efforts should be made to improve the models to classify ASD and TD scan-

Table 4. Frequency distribution of per image classification accuracy on the validation dataset

Range	Frequency
$[0, 0.5]$	1
$(0.5, 0.6]$	6
$(0.6, 0.7]$	17
$(0.7, 0.8]$	24
$(0.8, 1.0]$	12

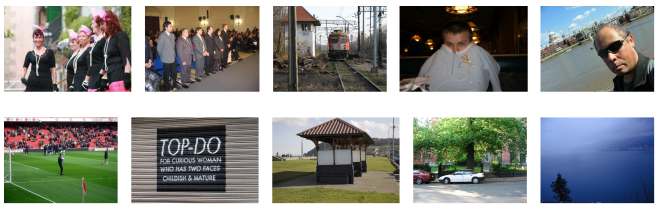

Fig. 3. Top-5 (first row) and bottom-5 (second row) images in the validation dataset

paths.

We further investigate the model performance for each image. Table 4 shows the frequency distribution of per image classification accuracy among all the images in the validation dataset, where each row shows the number of images having accuracy within the range. The top-5 and bottom-5 images are selected and depicted in Figure 3. The top-5 images are listed in the first row of the figure, while the bottom-5 are in the second row. Figure 3 shows that when there is a person, especially when the person is looking at the camera when the image was taken, the model performs better to classify ASD children from TD. When the images are natural scene or no face is shown in the images, the model performs worse.

5. CONCLUSION

In this paper, the SP-ASDNet framework for ASD classification based on the observer's gaze scanpaths is proposed. The CNN-LSTM architecture is adopted to extract the features from the saliency maps and to handle the sequence of the fixation locations in the scanpath. The proposed model achieves 74.22% accuracy on the validation dataset. In the future, the attentive mechanism will be investigated and integrated into the model to further improve the performance. We would investigate whether other image features can be incorporated in our proposed models to improve the performance as well. A thorough investigation on the differences in the scanpaths between ASD and TD children will also be conducted.

6. REFERENCES

[1] Jon Baio, Lisa Wiggins, Deborah L. Christensen, Matthew J. Maenner, Julie Daniels, Zachary Warren, Margaret Kurzius-Spencer, Walter Zahorodny, Cordelia Robinson Rosenberg, Tiffany White, Maureen S. Durkin, Pamela Imm, Loizos Nikolaou, Marshalyn Yeargin-Allsopp, Li-Ching Lee, Rebecca Harrington, Maya Lopez, Robert T. Fitzgerald, Amy Hewitt, Sydney Pettygrove, John N. Constantino, Alison Vehorn, Josephine Shenouda, Jennifer Hall-Lande, Kim Van Naarden Braun, , and Nicole F. Dowling, "Prevalence of autism spectrum disorder among children aged 8 years—autism and developmental disabilities monitoring network, 11 sites, united states, 2014," *MMWR Surveillance Summaries*, vol. 67, no. 6, pp. 1–23, April 2018.

[2] Xue Yang, Mei-Ling Shyu, Han-Qi Yu, Shi-Ming Sun, Nian-Sheng Yin, and Wei Chen, "Integrating image and textual information in human-robot interactions for children with autism spectrum disorder (ASD)," *IEEE Transactions on Multimedia*, vol. 21, no. 3, pp. 746–759, 2019.

[3] Peter C. Pantelis and Daniel P. Kennedy, "Deconstructing atypical eye gaze perception in autism spectrum disorder," *Scientific reports*, vol. 7, no. 14990, pp. 1–10, 2017.

[4] H. Duan, G. Zhai, X. Min, Z. Che, Y. Fang, X. Yang, J. Gutiérrez, and P. Le Callet, "A dataset of eye movements for the children with autism spectrum disorder," in *ACM Multimedia Systems Conference*, Amherst, MA, USA, June 2019.

[5] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[6] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao, "SALICON: saliency in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA, June 2015, pp. 1072–1080.

[7] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar, "A survey on deep learning: Algorithms, techniques, and applications," *ACM Computing Surveys*, vol. 51, no. 5, pp. 92:1–92:36, 2018.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1106–1114.

[9] Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu, "Multimodal deep representation learning for video classification," *World Wide Web*, pp. 1–17, 2018.

[10] Nian Liu and Junwei Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transaction on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.

[11] Ali Borji, "Saliency prediction in the deep learning era: An empirical investigation," *CoRR*, vol. abs/1810.03716, 2018.

[12] Sen Jia, "EML-NET: an expandable multi-layer network for saliency prediction," *CoRR*, vol. abs/1805.01047, pp. 1–10, 2018.

[13] Junting Pan, Cristian Canton-Ferrer, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *CoRR*, vol. abs/1701.01081, 2017.

[14] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE Transaction on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[15] James C. McPartland, Sara Jane Webb, Brandon Keehn, and Geraldine Dawson, "Patterns of visual attention to faces and objects in autism spectrum disorder," *Journal of autism and developmental disorders*, vol. 41, no. 2, pp. 148–157, 2011.

[16] Shuo Wang, Ming Jiang, Xavier Morin Duchesne, Elizabeth A. Laugeson, Daniel P. Kennedy, Ralph Adolphs, and Qi Zhao, "Atypical visual saliency in autism spectrum disorder quantified through model-based eye tracking," *Neuron*, vol. 88, no. 3, pp. 604–616, 2015.

[17] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yi Fang, Zhaohui Che, Xiaokang Yang, Cheng Zhi, Hua Yang, and Ning Liu, "Learning to predict where the children with asd look," in *IEEE International Conference on Image Processing*, 2018, pp. 704–708.

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

[19] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, Lille, France, July 2015, pp. 448–456.

[20] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.