A Comparative Study on Autism Spectrum Disorder Detection via 3D Convolutional Neural Networks

Kaijie Zhang[1,2], Wei Wang[1,2,3(✉)], Yijun Guo[2], Caifeng Shan[4,5], and Liang Wang[1,2,3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[2] Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition, Beijing, China
{kaijie.zhang,yijun.guo}@cripac.ia.ac.cn,
{wangwei,wangliang}@nlpr.ia.ac.cn

[3] Center for Excellence in Brain Science and Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing, China

[4] College of Electrical Engineering and Automation, Shandong University of Science and Technology, Qingdao, China
caifeng.shan@gmail.com

[5] Artificial Intelligence Research, Chinese Academy of Sciences, Beijing, China

Abstract. The prevalence of Autism Spectrum Disorder (ASD) in the United States has increased by 178% from 2000 to 2016. However, due to the lack of well-trained specialists and the time-consuming diagnostic process, many children are not able to be promptly diagnosed. Recently, several research have taken steps to explore automatic video-based ASD detection systems with the help of machine learning and deep learning models, such as support vector machine (SVM) and long short-term memory (LSTM) model. However, the models mentioned above could not extract effective features directly from raw videos. In this study, we aim to take advantages of 3D convolution-based deep learning models to aid video-based ASD detection. We explore three representative 3D convolutional neural networks (CNNs), including C3D, I3D and 3D ResNet. In addition, a new 3D convolutional model, called 3D ResNeSt, is also proposed based on ResNeSt. We evaluate these models on an ASD detection dataset. The experimental results show that, on average, all of the four 3D convolutional models can obtain competitive results when compared to the baseline using LSTM model. Our proposed 3D ResNeSt model achieves the best performance, which improves the average detection accuracy from 0.72 to 0.85.

Keywords: ASD detection · 3D convolution · 3D ResNeSt

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12661, pp. 546–558, 2021.
https://doi.org/10.1007/978-3-030-68763-2_42

1 Introduction

Autism Spectrum Disorder (ASD) is a developmental disorder, which could impair communication abilities and cause psychological and physical abnormalities. Recent research has shown that the prevalence of ASD was 18.5 per 1,000 (1 in 54) children aged 8 years across all 11 sites of the United States in 2016, while the prevalence was 6.7 per 1000 (1 in 150) in 2000, which indicates that the prevalence has increased by 178% in 16 years [12]. However, due to the lack of well-trained specialists and the time-consuming diagnostic process, many children cannot be diagnosed as early as possible. It is essential for children with ASD to receive early diagnosis since the importance of timely treatment for this kind of disease.

Machine learning and deep learning methods have achieved remarkable progress in many areas, such as image classification [4,10,15,16] and action recognition [1,3,8,13,19,20]. Recently, several research have taken steps to explore automatic video-based ASD detection systems with the help of these methods. Tariq et al. [17] adopted support vector machine (SVM) and logistic regression (LR) to identify possible ASD subjects by feeding them behavioral features assessed by non-expert raters from home videos. Zunino et al. [23] proposed an automated objective method using LSTM [5] model to discriminate between ASD and typically developing (TD) subjects. However, these models could not accept raw videos directly to extract effective features from both spatial and temporal dimensions.

In this study, we aim to take advantages of deep learning models to aid video-based ASD detection and achieve higher detection accuracy. In particular, we mainly consider 3D convolutional models, which can accept raw videos as input to extract effective features from both spatial and temporal dimensions. In our study, we explore three representative 3D convolutional neural networks, including C3D [19], 3D ResNet [3] and I3D [1]. In addition, we also propose a new 3D convolutional model by inflating all the 2D convolution and pooling kernels in the ResNeSt [22] model into 3D kernels, which is called 3D ResNeSt. These models are evaluted on the ASD detection dataset proposed in [23], which contains video clips of 40 subjects performing reach-to-grasp action with four different intentions, and to the best of our knowledge, this is the only publicly available video-based ASD detection dataset. Similar to [23], we adopt leave-one-out cross-validation strategy to evaluate and compare the performance of these 3D convolutional models. Our experimental results show that, on average, all of the four 3D convolutional models can achieve higher accuracy than [23], which means 3D convolutional models are indeed more suitable for this video-based ASD detection task. Our proposed 3D ResNeSt model outperforms the other three 3D CNNs when considering accuracy, f1 score and AUC. The average detection accuracy is improved from 0.72 to 0.85.

In summary, this paper has two major contributions: (1) We explore three representative 3D CNNs for ASD detection and the experimental results show that 3D convolutional models are more suitable for video-based ASD detection task. (2) A new 3D convolutional model is proposed based on ResNeSt, which

achieves the best performance on the ASD detection dataset proposed in [23]. The average detection accuracy is improved from 0.72 to 0.85.

2 Related Work

2.1 ASD Detection

The conventional diagnostic process of ASD needs well-trained specialists and it is also time-consuming. To reduce dependence on well-trained specialists, Tariq et al. [17] adopted machine learning models to identify possible ASD subjects by feeding them behavioral features assessed by non-expert raters from home videos. Zunino et al. [23] applied LSTM network to process video clips of children performing the same action to discriminate between ASD and TD subjects. Different from the methods mentioned above, Tian et al. [18] proposed a model called Temporal Pyramid Network to detect ASD typical actions and determine if repetitive behaviors appeared in videos to identify ASD and TD children. Liang et al. [11] proposed an unsupervised online learning model for ASD classification, which makes the classification system more scalable. Sun et al. [14] proposed a spatial attentional bilinear 3D convolutional network with LSTM model for fine-grained video analysis, which has achieved significant improvement on one class of the ASD detection dataset proposed in [23]. Besides video-based ASD detection methods, visual attention data are also proved to be able to provide effective features for ASD detection. Jiang et al. [9] analyzed the difference of eye fixations between ASD and TD subjects when viewing images and adopted deep learning model to extract features to distinguish between ASD and TD subjects. Chen et al. [2] presented a novel framework for automated and quantitative screening of ASD, a photo-taking task was introduced that subjects were asked to freely explore the environment and take some photos of the scene they are interested in. Then these photos were combined with the data collected in image-viewing task to train the ASD screening models.

In this study, we mainly focus on 3D convolutional models for video-based ASD detection task.

2.2 3D Convolutional Neural Networks (CNN)

3D CNN model was proposed for action recognition tasks, which can accept raw videos as input and has the advantage of extracting effective features from both spatial and temporal dimensions [8]. Based on 3D convolution operation, Tran et al. [19] proposed a deep 3D CNN model called C3D, which contains eight 3D convolution, five max-pooling, two fully connected layers and a softmax output layer, and it has achieved impressive results on action recognition tasks. The inception architecture was introduced in [15], which adopts 1×1 convolutions to reduce the parameters of neural networks without a significant performance penalty. Based on [15], Carreira et al. [1] proposed a new 3D convolutional model called Inflated 3D ConvNet (I3D) by inflating all the filters and pooling kernels

of the model proposed in [15]. Residual learning framework has been proved to be able to achieve excellent performance when training very deep neural networks [4]. Inspired by [4], Hara et al. [3] proposed a deeper 3D convolutional model by changing the 2D convolution operation of the ResNet [4] model into 3D convolution operation and achieved better performance than relatively shallow networks.

In our experiments, we explore three 3D convolutional neural networks, including C3D [19], 3D ResNet [3] and I3D [1]. In addition, we also propose a new 3D convolutional model based on ResNeSt [22]. The structure and implementation details of these models will be shown in Sect. 3 and 4.

3 3D CNN for ASD

The abnormal behaviors linked to ASD can be well recorded in videos. As 3D CNNs can accept raw videos as input and have the advantage of extracting effective features from both spatial and temporal dimensions, it is a good choice to apply 3D CNNs for video-based ASD detection task. The overall detection procedure of our method is shown in Fig. 1. As depicted in Fig. 1, when given a video, Gaussian smoothing is applied to each frame first, and then we will randomly sample 16 consecutive frames n times from the video. Specifically, during the training process, n is 1, and during the validation process, n is 10. These clips are then fed to 3D convolutional models. The outputs of these clips will be averaged to form the final result. In this study, we consider three representative 3D CNNs in our experiments, including C3D [19], 3D ResNet [3] and I3D [1]. Recently, a new variant of ResNet model, called ResNeSt, was proposed in [22], which achieves impressive results in object detection, instance segmentation and semantic segmentation tasks. Based on ResNeSt model, we propose a new 3D convolutional model called 3D ResNeSt. We will introduce the detailed structure of these models in the following subsections.

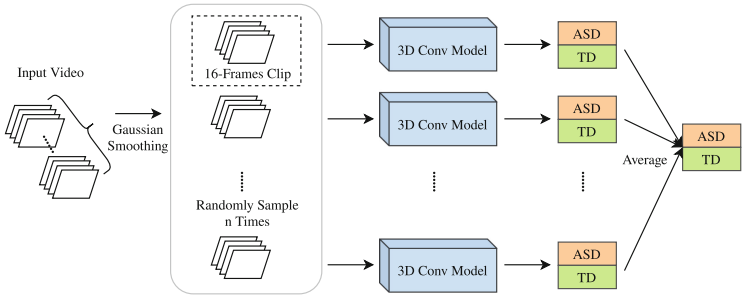

Fig. 1. The detection procedure of our method.

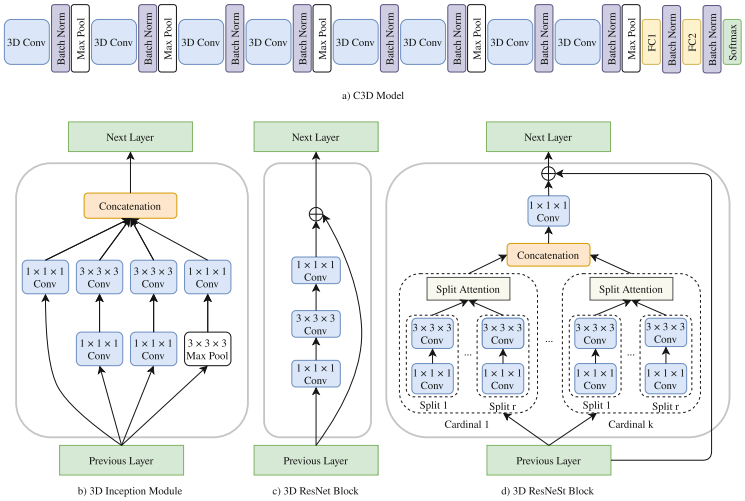

Fig. 2. The structure of the models used in our experiments. a) C3D model, b) 3D Inception module, c) 3D ResNet block, d) 3D ResNeSt block of our proposed model, all kernels of the convolution and the polling operation in the ResNeSt [22] block are inflated from N × N to N × N × N

3.1 C3D

C3D model was proposed in [19], which contains eight 3D convolution, five max-pooling, two fully connected layers and a softmax output layer, and it has achieved impressive results on action recognition tasks. In our experiments, we make a small change to the original model. A batch normalization layer is added after all the convolution and fully connected layers, which could solve the internal covariate shift problem and accelerate the training process [7]. The overview of the C3D model structure is shown in Fig. 2. a), and the batch normalization layer is added explicitly in Fig. 2. a) to distinguish from the original C3D model.

3.2 I3D

Compared to 2D CNNs, 3D CNNs always have more parameters due to the additional kernel dimension, which makes the model harder to train. To handle this problem, Carreira et al. [1] proposed the I3D model by inflating all the filters and pooling kernels of the model proposed in [15]. Figure 2. b) shows the crucial architecture of I3D called 3D Inception Module, $1 \times 1 \times 1$ convolutions are applied before the expensive $3 \times 3 \times 3$ convolutions, which reduces the number of parameters and allows the model to increase in both width and depth without getting into computational difficulties. In our paper, as the size of the input frames are set to $16 \times 112 \times 112$ (time × height × width), we implemented a small variation of I3D model, the kernel size of the first convolution layer is changed from $7 \times 7 \times 7$ to $3 \times 3 \times 3$ with a stride of $2 \times 1 \times 1$, the kernel

size of the last average pooling layer is changed from $2 \times 7 \times 7$ to $2 \times 5 \times 5$ with a stride of $1 \times 1 \times 1$, and the last convolution layer is replaced with 2 fully connected layers to get the final outputs.

3.3 3D ResNet

Residual learning framework, which introduces shortcut connections that bypass a signal from one layer to the next, has been proved to be able to achieve excellent performance when training very deep neural networks [4]. Inspired by [4], Hara et al. [3] proposed a deeper 3D convolutional model by changing the 2D convolution operation of the ResNet [4] model into 3D convolution operation and achieved better performance than relatively shallow networks. The 3D ResNet model has various versions with different total layers. Due to the time consuming training process, in this study, we only consider the 50-layers model. Figure 2. c) shows the residual block of the 50-layers 3D ResNet model, which contains three convolution layers, including two $1 \times 1 \times 1$ and one $3 \times 3 \times 3$ convolution layers.

3.4 The Proposed 3D ResNeSt

Multi-path representation, group convolution and channel-attention mechanism have been proved to be successful in many computer vision tasks [6,15,21]. Inspired by these methods, Zhang et al. [22] proposed the ResNeSt model, which generalizes the channel-wise attention into feature-map group representation. Based on ResNeSt, we propose a new 3D convolutional model called 3D ResNeSt. Figure 2. d) shows the key block of the proposed model. Our model preserves the structure of ResNeSt. Outputs from previous layer are divided into several cardinal groups and finer-grained splits when fed to the 3D ResNeSt blocks. We also adopt the split attention operation to aggregate all the splits in each cardinal group like [22]. In our experiments, the number of cardinal groups and splits are set to 1 and 2 respectively, which has been proved to be a good trade-off between speed, accuracy and memory usage in [22]. All the kernels of the convolution and the polling operation are inflated from $N \times N$ to $N \times N \times N$, except the kernel of the optional average pooling layer, the kernel of this layer is inflated from 3×3 to $1 \times 3 \times 3$. To ensure comparability, we also only consider 50-layers 3D ResNeSt model in this study.

3.5 Dataset

The ASD detection dataset [23] used in our experiments is downloaded from https://pavis.iit.it/datasets/autism-spectrum-disorder-detection-dataset with the authors' authorization. This dataset contains video clips of 40 subjects performing reach-to-grasp action with four different intentions. Among the 40 subjects, 20 are ASD children without accompanying intellectual impairment and 20 are TD children. The reach-to-grasp action mentioned above refers

Fig. 3. Sample frames from the dataset after applying Gaussian smoothing.

to grasping an object (a bottle), all subjects are asked to perform the same action with four intentions, including 1) to place it into a box (grasp-to-place), 2) to pour some water into a glass (grasp-to-pour), 3) to pass the bottle to a co-actor, who would then place the bottle into the box (pass-to-place), 4) to pass the bottle to a co-actor, who would then pour some water (pass-to-pour). And for short, we use class1, class2, class3, class4 to represent the above four classes of reach-to-grasp actions respectively.

3.6 Data Preprocessing

Similar to [23], we apply Gaussian smoothing over all the frames to reduce details of visual appearance. The original resolution of the video frames is 1280 × 720 pixels. However, some frames may contain subjects' head or body. In order to remove this information, the right part of all the frames are cropped a width of 150 pixels, and then the resolution of the remaining parts of these frames becomes 1130 × 720 pixels. Figure 3 shows some sample frames from the dataset after applying Gaussian smoothing.

Table 1. Setup of learning rate and learning rate decay of the four 3D convolutional models.

	C3D	I3D	3D ResNet	3D ResNeSt
Base learning rate	0.0001			
Epoch of learning rate decay	80	90	100	90

Table 2. Performance of the models used in our experiments and the LSTM model used in [23] evaluated on the ASD detection dataset.

		[23]	C3D	I3D	3D ResNet	3D ResNeSt
Class1	Accuracy	0.67	0.69	0.74	**0.79**	0.77
	F1	0.65	0.67	0.74	**0.80**	0.74
	Sensitivity	0.63	0.63	0.74	**0.84**	0.68
	Specificity	0.70	0.75	0.75	0.75	**0.85**
	AUC	0.74	0.82	0.79	**0.84**	0.79
Class2	Accuracy	0.77	0.72	0.72	0.69	**0.79**
	F1	**0.77**	0.69	0.69	0.68	0.76
	Sensitivity	**0.79**	0.63	0.63	0.68	0.68
	Specificity	0.75	0.80	0.80	0.70	**0.90**
	AUC	**0.86**	0.79	0.80	0.79	0.84
Class3	Accuracy	0.69	0.79	**0.85**	0.79	**0.85**
	F1	0.67	0.79	0.83	0.79	**0.82**
	Sensitivity	0.63	**0.79**	**0.79**	**0.79**	0.74
	Specificity	0.75	0.80	0.90	0.80	**0.95**
	AUC	0.76	0.89	0.88	0.86	**0.91**
Class4	Accuracy	0.59	0.77	0.74	0.77	**0.82**
	F1	0.53	0.71	0.74	0.78	**0.81**
	Sensitivity	0.47	0.58	0.74	**0.84**	0.79
	Specificity	0.70	**0.95**	0.75	0.70	0.85
	AUC	0.75	0.84	0.87	0.86	**0.90**
Average	Accuracy	0.72	0.79	0.79	0.79	**0.85**
	F1	0.70	0.76	0.78	0.80	**0.82**
	Sensitivity	0.68	0.68	0.74	**0.8**4	0.74
	Specificity	0.75	0.90	0.85	0.75	**0.95**
	AUC	0.84	0.87	0.88	0.86	**0.90**

4 Experiments and Results

4.1 Implementation Details

As the ASD detection dataset only contains 40 subjects, for better evaluating and comparing the performance of these 3D convolutional models, we also adopt leave-one-out cross-validation strategy like [23], which is more challenging than the usual cross-validation strategy. The four classes of videos in the ASD detection dataset are processed separately, which means we will perform leave-one-out cross-validation procedure on the four different classes of videos respectively. This leads to a total number of 160 models to be trained when evaluate one 3D Convolutional model on this dataset, which is very time consuming. All of our models are trained on a single GPU. We adopt adaptive moment estimation (Adam) to optimize our model. The cropped frames are resized to 112 $\times$ 112 pixels to form the input. The base learning rate of these models is all 0.0001 and is divided by 10 after specific epochs (shown in Table 1).

Table 3. Leave-one-out cross-validation results compared with the results reported in [23]. In the average column, probabilities greater than 0.5 are highlighted in bold.

		Class1		Class2		Class3		Class4		Average	
		[23]	3D ResNeSt	[23]	3D ResNeSt	[23]	3D ResNeSt	[23]	3D ResNeSt	[23]	3D ResNeSt
ASD	1	0.67	0.68	0.80	1.00	0.73	0.49	0.27	0.05	**0.62**	**0.55**
	2	0.09	0.94	1.00	1.00	0.08	0.75	0.92	0.98	**0.52**	**0.92**
	3	–	–	–	–	–	–	–	–	–	–
	4	0.83	0.72	0.64	0.52	0.00	0.92	0.08	0.01	0.39	**0.54**
	5	0.17	0.71	0.67	0.24	1.00	0.93	0.45	0.92	**0.57**	**0.70**
	6	0.00	0.29	1.00	0.71	0.08	0.67	0.08	0.97	0.29	**0.66**
	7	0.25	0.32	0.00	0.49	0.83	0.36	0.08	0.22	0.29	0.35
	8	0.08	0.01	0.08	0.00	0.42	0.00	0.25	0.68	0.21	0.17
	9	0.92	1.00	1.00	0.99	1.00	1.00	1.00	1.00	**0.98**	**1.00**
	10	0.92	0.99	1.00	0.98	0.58	1.00	1.00	1.00	**0.88**	**0.99**
	11	1.00	1.00	0.92	1.00	1.00	1.00	1.00	0.99	**0.98**	**1.00**
	12	1.00	0.98	1.00	0.99	0.92	1.00	0.42	1.00	**0.84**	**0.99**
	13	1.00	0.94	0.83	0.98	0.83	0.90	1.00	1.00	**0.92**	**0.96**
	14	1.00	1.00	0.75	1.00	1.00	1.00	0.50	1.00	**0.81**	**1.00**
	15	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**1.00**	**1.00**
	16	0.33	0.07	0.42	0.27	0.17	0.12	0.45	0.57	0.34	0.26
	17	0.17	0.06	0.00	0.16	0.09	0.81	0.17	0.75	0.11	0.44
	18	1.00	0.03	1.00	0.13	0.83	0.44	1.00	0.38	**0.96**	0.24
	19	0.91	1.00	1.00	1.00	0.50	1.00	1.00	1.00	**0.85**	**1.00**
	20	0.75	1.00	1.00	1.00	1.00	1.00	0.92	1.00	**0.92**	**1.00**
TD	21	0.90	1.00	0.75	0.76	0.55	0.56	0.58	0.84	**0.70**	**0.79**
	22	0.33	0.92	0.50	0.69	0.67	0.53	0.08	0.56	0.40	**0.67**
	23	0.42	0.56	0.75	0.67	0.00	0.64	0.00	0.62	0.29	**0.62**
	24	0.58	0.53	0.40	0.82	0.36	0.40	0.42	0.97	0.44	**0.68**
	25	0.64	0.67	0.92	0.92	0.83	0.88	0.50	0.93	**0.72**	**0.85**
	26	1.00	0.76	1.00	0.79	1.00	0.75	0.83	0.44	**0.96**	**0.68**
	27	0.91	0.83	0.75	0.94	1.00	0.70	1.00	0.42	**0.92**	**0.72**
	28	0.67	0.23	1.00	0.57	1.00	0.89	0.92	0.65	**0.90**	**0.59**
	29	0.67	1.00	0.33	1.00	1.00	1.00	0.92	1.00	**0.73**	**1.00**
	30	1.00	1.00	0.75	1.00	0.83	1.00	0.50	1.00	**0.77**	**1.00**
	31	0.92	0.96	0.83	0.94	1.00	0.99	0.92	1.00	**0.92**	**0.97**
	32	0.75	0.98	0.92	0.96	0.91	1.00	1.00	0.99	**0.90**	**0.98**
	33	0.33	1.00	1.00	1.00	0.92	1.00	0.75	1.00	**0.75**	**1.00**
	34	0.50	1.00	1.00	1.00	1.00	1.00	1.00	1.00	**0.88**	**1.00**
	35	1.00	1.00	1.00	1.00	1.00	0.98	1.00	1.00	**1.00**	**0.99**
	36	0.45	0.00	0.25	0.00	0.08	0.51	0.67	0.04	0.36	0.14
	37	0.82	0.82	1.00	0.78	0.82	0.64	0.20	0.75	**0.71**	**0.75**
	38	0.18	0.89	0.50	0.83	0.36	0.98	0.83	0.97	0.47	**0.92**
	39	0.64	0.23	0.67	0.91	0.45	0.96	1.00	0.98	**0.69**	**0.77**
	40	1.00	0.56	1.00	0.07	0.83	0.95	1.00	0.95	**0.96**	**0.63**
Accuracy ($p > 0.5$)		0.67	0.77	0.77	0.79	0.69	0.85	0.59	0.82	0.72	0.85

4.2 Experimental Results

As mentioned above, in our experiments, the four classes of videos in the ASD detection dataset are separately processed and these 3D convolutional models perform leave-one-out cross-validation on the four classes of videos respectively. In this paper, we mainly consider the average performance on all classes, and as

[14] only report the results of one class, we did not include [14] in our comparison. As for the average performance, [23] is currently state-of-the-art method. Therefore, we mainly compare our results with [23]. We use detection accuracy, f1 score, sensitivity, specificity, and area under the receiver operating characteristic curve (AUROC/AUC) as the metrics to evaluate our models. We report the ASD detection performances measured by these metrics in Table 2. As [23] did not contain the results of these metrics, we use the data reported in [23] to calculate them. As for the average results of these metrics, we first calculate the average ASD and TD probabilities of the leave-one-out cross-validation results on the four classes, and then use the average probabilities to calculate these metrics. From Table 2, we can find that, on average, all of the four 3D convolutional models can achieve higher accuracy and f1 score than [23], which means 3D convolutional models are indeed more suitable for this task. Our proposed 3D ResNeSt model achieves the best performance when considering accuracy, f1 score and AUC and the average detection accuracy is improved from 0.72 to 0.85. If we explore the results class by class, we can draw the following conclusions. For class1, the 3D ResNet model is the best choice among the five models, which acquires the highest f1 score, accuracy and AUC. For class2, [23] may be more suitable for this class of videos, as it achieves the highest f1 score and AUC, although the 3D ResNeSt model achieves higher accuracy. For class3 and class4, our proposed model is the best choice among the five models, which achieves much better performance than [23]. The detection accuracy of class3 is improved from 0.69 to 0.85 and class4 is improved from 0.59 to 0.82.

In order to make a more detailed comparison between our proposed 3D ResNeSt model that achieves the best performance in our experiments and the published research in [23], we report the leave-one-out cross-validation results in Table 3. Each line in Table 3 refers to the outputs of a different subject left out. The first half (subject 1–20) represents the results with ASD subjects left out and the rest with TD ones left out (subject 21–40). For subject 1–20, the value in cell represents the probability of being predicted as ASD, and for subject 21–40, it represents the probability of being predicted as TD. From Table 3, we can find that, on average, [23] and our model both achieve better performance in TD group than in ASD group, which means both achieve higher specificity than sensitivity. Similar to [23], some subjects could be perfectly classified with our model in all classes of the ASD detection dataset, like subject 14, 15, 19 and 20 in the ASD group and subject 29, 30, 33 and 34 in the TD group. However, for some subjects, our model could not classify them in any class of the dataset, like subject 7 and 18 in the ASD group. Globally, our proposed 3D ResNeSt model achieves much better performance than [23].

To further evaluate the performance of our model, we report the average detection accuracy of the 3D ResNeSt model and the LSTM model used in [23] when the threshold varies from 0.5 to 0.95 with a step of 0.05 in Fig. 4. We can find that when the threshold is greater than 0.7, our model is much more robust than [23], and even when the threshold is 0.95, our model still can achieve an accuracy of 0.41, while the accuracy of LSTM model has reduced to 0.18.

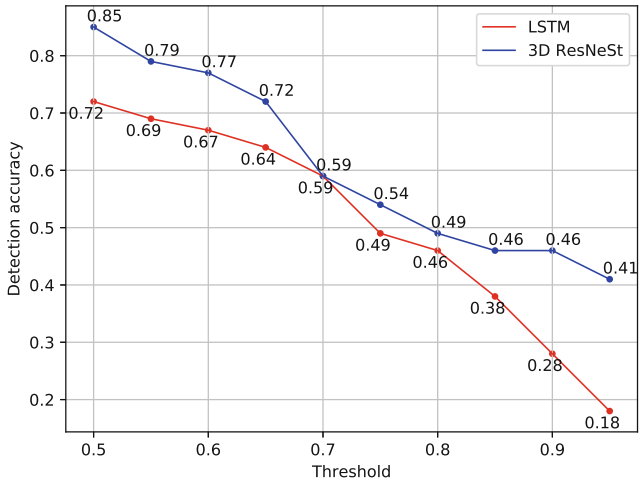

Fig. 4. Average detection accuracy acquired by 3D ResNeSt and LSTM [23] with different threshold.

5 Conclusion

In this paper, we explore three representative 3D CNNs for ASD detection task, including C3D, I3D and 3D ResNet, our experimental results show that 3D convolutional models are more suitable for video-based ASD detection task. And we also proposed a new 3D convolutional model based on ResNeSt, which achieves the best performance on the ASD detection dataset reported in [23]. However, when compared to other action recognition tasks, the ASD detection dataset is still too small. If we want better evaluation of these models, we still need larger datasets. Meanwhile, we should also take computational efficiency into consideration when video datasets become larger. In addition, we did not explore how these models make predictions, as for the future work, studying on how deep learning models make predictions may provide more meaningful information for diagnosis of ASD.

Acknowledgments. This work is jointly supported by the National Natural Science Foundation of China (NO. 61976214, 61972188), and Shandong Provincial Key Research and Development Program (Major Scientific and Technological Innovation Project) (NO. 2019JZZY010119).

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)

2. Chen, S., Zhao, Q.: Attention-based autism spectrum disorder screening with privileged modality. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1181–1190 (2019)
3. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 3154–3160 (2017)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
5. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
7. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
8. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2012)
9. Jiang, M., Zhao, Q.: Learning visual attention to identify people with autism spectrum disorder. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3267–3276 (2017)
10. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
11. Liang, S., Loo, C.K., Md Sabri, A.Q.: Autism spectrum disorder classification in videos: a hybrid of temporal coherency deep networks and self-organizing dual memory approach. In: Kim, K.J., Kim, H.-Y. (eds.) Information Science and Applications. LNEE, vol. 621, pp. 421–430. Springer, Singapore (2020). https://doi.org/10.1007/978-981-15-1465-4_42
12. Maenner, M.J., Shaw, K.A., Baio, J., et al.: Prevalence of autism spectrum disorder among children aged 8 years-autism and developmental disabilities monitoring network, 11 sites, united states, 2016. MMWR Surveill. Summ. **69**(4), 1 (2020)
13. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
14. Sun, K., Li, L., Li, L., He, N., Zhu, J.: Spatial attentional bilinear 3d convolutional network for video-based autism spectrum disorder detection. In: ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3387–3391. IEEE (2020)
15. Szegedy, C., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
16. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818–2826 (2016)
17. Tariq, Q., Daniels, J., Schwartz, J.N., Washington, P., Kalantarian, H., Wall, D.P.: Mobile detection of autism through machine learning on home video: a development and prospective validation study. PLoS Med. **15**(11), e1002705 (2018)
18. Tian, Y., Min, X., Zhai, G., Gao, Z.: Video-based early asd detection via temporal pyramid networks. In: 2019 IEEE International Conference on Multimedia and Expo (ICME), pp. 272–277. IEEE (2019)

19. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
20. Wang, L., et al.: Temporal segment networks: towards good practices for deep action recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9912, pp. 20–36. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46484-8_2
21. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1492–1500 (2017)
22. Zhang, H., et al.: ResNeSt: split-attention networks. arXiv preprint arXiv:2004.08955 (2020)
23. Zunino, A., et al.: Video gesture analysis for autism spectrum disorder detection. In: 2018 24th International Conference on Pattern Recognition (ICPR), pp. 3421–3426. IEEE (2018)