SPATIAL ATTENTIONAL BILINEAR 3D CONVOLUTIONAL NETWORK FOR VIDEO-BASED AUTISM SPECTRUM DISORDER DETECTION

Kangbo Sun[†] *Lin Li*[†] *Lianqiang Li*[†] *Ningyu He*[†] *Jie Zhu*[†⋆]

[†] Department of Electronic Engineering, Shanghai Jiaotong University, China
Email:{kangbosun, lilin7, sjtu_llq, ruby_yu, zhujie}@sjtu.edu.cn

ABSTRACT

Video-based Autism Spectrum Disorder (ASD) detection is a challenge to most video classification networks due to the high degree of similarity between categories. Bilinear pooling is a second-order method, which is widely used in fine-grained visual recognition. However, the average summation in bilinear pooling limits its ability to perceive spatial information, which is detrimental to fine-grained visual recognition. In this paper, we propose spatial attentional bilinear pooling to enhance its spatial information extraction without significantly increasing the parameters. Further, we propose a fine-grained action recognition network named SA-B3D with LSTM model for video-based ASD detection. The proposed model can focus on more discriminative regions dynamically and effectively. Compared with state-of-the-art models, the proposed model achieves significant improvement on video-based ASD dataset.

Index Terms— Fine-grained, Attention mechanism, Bilinear pooling, Autism diagnosis, Action Detection

1. INTRODUCTION

Detection of Autism Spectrum Disorder (ASD) is an important step in the treatment of children with ASD. Video-based detection of ASD is a challenge to the most mainstream neural networks. The actions of Typically Developing (TD) children and ASD children have analogous patterns, which are difficult to be distinguished. Without specific knowledge, it could be a challenging task even for human beings.

Traditional video analysis methods [1, 2] described video content by hand-crafted feature descriptors. Currently, inspired by the successes of deep networks in image recognition, C3D [3] and two-stream convolutional networks [4] were proposed for capturing spatiotemporal representation of video action. However, video frames are too many for both C3D and two-stream convolutional networks. Most architectures [3, 4, 5, 6] try to reduce the input video frames by subsampling or splitting integer videos into short snippets about 16-frame clips, which may lose essential action details. In the video-based ASD detection, essential information exists only on a small number of consecutive frames, and the subsampling operation will greatly reduce the performance of the classification. In our work, a video-level model for ASD detection is proposed, which is shown in Fig. 1. To classify ASD videos in video level, our proposed model utilizes an LSTM network to extract temporal information on consecutive features produced by the C3D network.

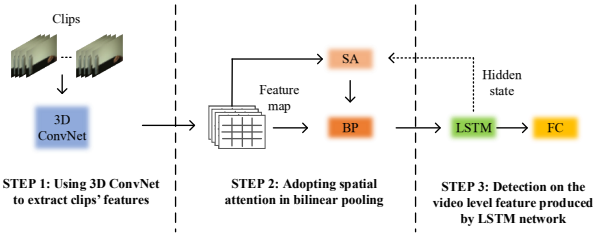

Fig. 1. Overview of our SA-B3D with LSTM model. The dashed line with an arrow indicates that the previous hidden state of the LSTM is a part of the input to the SA module.

Bilinear pooling method has achieved excellent performance [7, 8, 9] in the fine-grained visual task. This method is designed to extract the second-order information of the features extracted by CNNs. The second-order information comes from mutual information between feature channels, which greatly increases the information for classification. Similarly, bilinear pooling can be adopted to enhance the features produced by the C3D network. However, the average summation of bilinear pool will cause the network to lose spatial information to a certain extent. In our work, we adopt an attention mechanism to build spatial attentional bilinear pooling, which, to a certain extent, overcomes the weakness of bilinear pooling in extracting spatial information without significantly increasing the parameters.

In this paper, a novel Spatial Attentional Bilinear 3D convolutional network (SA-B3D) with LSTM model (Fig. 1) is proposed for fine-grained video analysis. Our contributions can be summarized as follows: (a) Video-level model is adopted for ASD detection, which outperforms frame-level or clip-level models. (b) Spatial attentional bilinear pooling is proposed to overcome the weakness of bilinear pooling in ex-

⋆ Corresponding author, Email: zhujie@sjtu.edu.cn

tracting spatial information. (c) Our proposed SA-B3D with LSTM model achieves state-of-the-art and achieves 87.17% detection accuracy on the ASD dataset [10].

The rest of this paper is organized as follows. Section 2 introduces the related works. Section 3 describes the proposed approach in detail. Section 4 provides the experiments and results. Finally, conclusions are drawn in Section 5.

2. RELATED WORKS

Autism spectrum disorder detection. Early detection methods based on ASD questionnaire exams like ADOS [11], magnetic resonance imaging or genetic engineer are time-consuming and require professional diagnoses, which means these methods cannot be widely used for early screening. Recently, Zunino et al. [10] proposed a grabbing bottle dataset performed by ASD and TD children. There is a difference in the performance of TD children and ASD children at the moment of grabbing a bottle. It means that we can classify whether actions are performed by a TD or an ASD child, by only processing the part of video data recording the grasping gesture [10]. Zunino et al. adopted the network proposed by Sharma et al. [12] to distinguish ASD children and TD children. Y. Tian et al. [13] proposed an action dataset named ASD40h, and they used a 3D convolutional network to extract temporal features. Our work is based on the grabbing bottle dataset [10]. Different from the clip-level model in [10], our proposed SA-B3D model is a video-level model.

Bilinear pooling. Lin et al. [7] adopted bilinear pooling before the fully connected layers and achieved remarkable improvements on fine-grained visual recognition. Many works about bilinear pooling focused on simplifying bilinear operations to reduce computational costs like [14] or designing various bilinear networks to improve performance like [15]. In our work, we propose weighted sum bilinear pooling and adopt a spatial attentional mechanism to enhance its ability to perceive spatial information without significantly increasing the parameters.

Spatial attentional mechanism. Attentional model was proposed by Xu et al. [16] for image captioning. Their soft-attention mechanism that averages the spatial features with attentional weights was widely adopted in visual captioning [16, 17] and question answering task [18, 19]. Chen et al. [20] applied spatial attentional mechanism to multiple layers of a CNN for image captioning. In our SA-B3D model, we adopt a soft-attention mechanism on the C3D features using the spatiotemporal features of the current clip and the hidden state of previous LSTM, which improves the detection accuracy significantly compared with the baseline model.

3. APPROACH

In this section, the weighted sum bilinear pooling and the proposed SA-B3D with LSTM model are introduced in detail.

3.1. Weighted sum bilinear pooling

Gao et al. [14] have proved that part of the reason for the success of bilinear pooling is its efficient extraction of the second-order information on channels. However, the summation way of the traditional bilinear pooling weakens its spatial information extraction capability. In this paper, we propose weighted sum bilinear pooling.

It is assumed that $F \in R^{S\times C}$ is the feature map obtained from the 3D convolutional network, where C is the number of channels, and $S = T * H * W$ is the size of feature map. Bilinear pooling is defined as:

$$B = F^T F \tag{1}$$

where $B \in R^{C\times C}$ is the output of bilinear pooling layer. For each element $B_{i,j}$ in B, there is:

$$B_{i,j} = {F_i}^T F_j = (X^i)^T X^j = \sum_{k=1}^{S} {x_k}^i {x_k}^j \tag{2}$$

where $B_{i,j}$ is the product of spatial location of the original feature in channel i and j, $X^i = ({x_1}^i, {x_2}^i, ..., {x_S}^i)^T$ means the feature on channel i, and ${x_j}^i$ means the feature on the position j of the channel i.

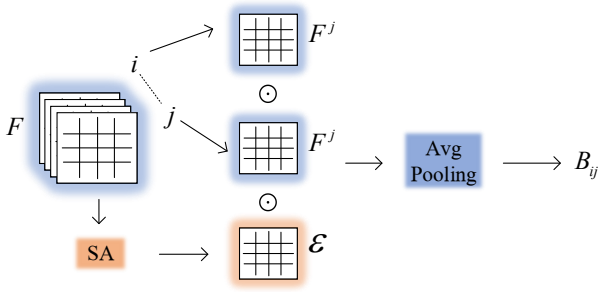

Fig. 2. Generation of spatial attentional bilinear feature B_{ij}

However, the method of sum pooling does not consider the imbalance of spatial information distribution. This method of treating spatial information equally cannot extract key information effectively. To characterize the effect of spatial position on pooling results, we introduce a weighting factor $\mathcal{E} \in R^{S\times C}$. As a result, Eq. 1 and Eq. 2 can be rewritten as:

$$B = (\mathcal{E} \odot F)^T (\mathcal{E} \odot F) \tag{3}$$

$$B_{i,j} = (\mathcal{E}^i \odot X^i)^T (\mathcal{E}^j \odot X^j) = \sum_{k=1}^{S} {\varepsilon_k}^{i,j} {x_k}^i {x_k}^j \tag{4}$$

where $\odot$ means matrix point multiplication, $\mathcal{E} \in R^{S\times C}$ means weight matrix and ${\varepsilon_k}^{i,j}$ equals ${\varepsilon_k}^i * {\varepsilon_k}^j$.

To enable the network to focus on the discriminative areas of the image and determine the distribution of weights, we utilize an attention mechanism to generate a learnable weight, which has same distribution on all channels.

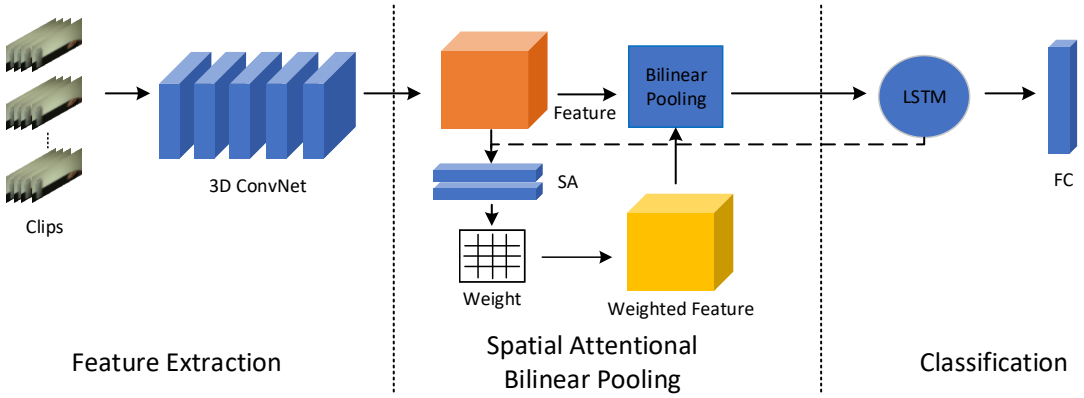

Fig. 3. Implementation of spatial attentional bilinear 3D convolutional network with LSTM. **3D ConvNet**: five 3D convolutional layers followed by four max-pooling layers and the first max-pooling layer does not pool the time dimension. **SA**: two fully connected layers with tanh and sigmoid activation. The dashed line with an arrow indicates that the previous hidden state of the LSTM is a part of the input to the SA module.

Therefore, the weighted sum bilinear pooling in Eq. 4 is shown as:

$$B_{i,j} = \sum_{k=1}^{S} \varepsilon_k x_k{}^i x_k{}^j \tag{5}$$

where ε_k is the k-th element in $\mathcal{E} \in R^S$. The calculation process of B_{ij} is shown in Fig. 2.

3.2. SA-B3D with LSTM model

In fine-grained video recognition task, discriminative information in video tends to exist in a small timing range. The preprocessing method of frame extraction may greatly hurt discriminative information. Besides, the increasing of parameters caused by 3D convolution limits the number of frames in each clip. To address this issue, we use the LSTM network to extract the information of descriptor produced by SA-B3D network with clips as input. A two-layer fully connected network is adopted to generate spatial weights that can be trained with the backpropagation algorithm in the end-to-end training. With the LSTM network, the weight matrix $\mathcal{E}$ generation formula is shown as:

$$\mathcal{A}_t = Tanh(W_{fh}^T(\mathcal{F}_t \oplus H_{t-1})) \tag{6}$$

$$\mathcal{E}_t = Sigmoid(W_a^T \mathcal{A}_t) \tag{7}$$

where the subscript t represents the t-th clip, H_{t-1} is the hidden state of LSTM cell at step $t-1$, $\mathcal{F}_t \in R^{SC}$ is descriptor produced by the C3D network, $W_{fh} \in R^{SC \times D}$ and $W_a \in R^{D \times S}$ are learnable weight matrices, $\mathcal{E}_t \in R^S$ is the spatial attention weight matrix, which has the same weight distribution on all channels. The implementation of spatial attentional bilinear 3D convolutional network with LSTM model is shown in Fig. 3.

4. EXPERIMENT

In this section, we introduce the experiments in detail. First, we introduce the ASD dataset used. Second, our details and specific implementation of the experiment are introduced. Third, we compare our methods with other state-of-the-art methods. Finally, we visualize the results of spatial attention.

4.1. Dataset

We conduct experiments on the video-based ASD dataset [10] that has extremely indistinguishable actions, and we evaluate our models on videos with condition pouring [10]. It has been shown that it is indeed possible to eliminate their ambiguity through video gestures with subtle differences. Same as Zunino et al., the proposers of the dataset, we select the one-subject-out testing procedure. We extract the frames of the video and use 16 consecutive frames as a clip. Since the right half of the frames do not contain the action of grabbing bottle, the frames are cropped to the left half of the original frame and resized to 128 * 128 in all experiments.

4.2. Model implementation

Implementation detail. We utilize a small 3D convolutional network with five 3D convolutional layers followed by max-pooling. The first max-pooling does not pool the time dimension. To reduce the dimension, we use 3D convolutional layers with a size of 32*1*1*1 to compress the number of channels to get the description of the clip. We train all models using Adam with a batch size of 4, an initial learning rate of 2×10^{-5}. We use the signed square root and L2 normalization after bilinear pooling, which is widely applied in [7, 14, 15]. The training of the proposed SA-B3D with LSTM model is divided into two parts. First, we use a cross-entropy loss function to pre-train modules other than the SA module. Then, we fine-tune the whole model with the loss function determined in Eq. 8.

Network	PA	Avg_acc%	Dec_acc%
C3D with LSTM	7.7M	71.97	71.79
B3D with LSTM	24.5M	78.44	79.48
SA-B3D with LSTM	25.6M	**80.90**	**84.61**
Zunino et al. [10]	—	75.50	76.92
C3D with LSTM	7.7M	72.68	79.48
B3D with LSTM	24.5M	77.73	82.05
SA-B3D with LSTM	25.6M	**82.56**	**87.17**

Table 1. Performance on the ASD dataset [10]

Loss function. To force the spatial attentional network to be more discriminative, an attention constraint is adopted to make the attention map have a larger variance. The loss function is shown in Eq. 8.

$$Loss = -\sum_{i=1}^{N} y_i log\hat{y}_i + \lambda(1 - \sum_{i=1}^{S} \frac{(a_i - \mu)^2}{S}) \quad (8)$$

where y_i is the one hot label vector, $\hat{y}_i$ is the vector of class probabilities, a_i and μ is i-th weight and the mean value of spatial attention map respectively, $\lambda = 0.001$ is the spatial attention constraint coefficient.

4.3. Performance

The comparison of the proposed SA-B3D model with other methods is shown in Table 1. We use average accuracy and detection accuracy to evaluate the performance of the model. The average accuracy means the proportion of the video samples that are correctly predicted and the detection accuracy means the proportion of the object samples that are correctly predicted. If more than half of the video samples are predicted correctly, the object samples are predicted correctly [10]. PA represents the parameter amount of the model. The top half of Table 1 shows the results of our five-fold testing method with the above configurations, and the five-fold trained models are used as pre-trained models for the one-subject-out testing method shown in the bottom half.

In the five-fold experiment, the proposed SA-B3D model outperforms C3D baseline with 8.93% and 12.82% improvement on average and detection accuracy, respectively. In the one-subject-out testing experiment, SA-B3D model outperforms C3D baseline with 9.88% and 7.69% improvement on average and detection accuracy, respectively, and outperforms B3D model with 4.83% and 5.12% improvement on average and detection accuracy, respectively. The proposed SA-B3D model achieves state-of-the-art with both evaluation methods, and the amount of SA-B3D parameters is only 4.5% more compared with B3D.

4.4. Visualization

The features captured by the C3D, B3D and SA-B3D network are visualized in Fig. 4. To ensure sufficient features for visualization, we selected the five-fold testing model for feature extraction. The features extracted by the SA-B3D model have larger inter-class distances and smaller intra-class distances, which means that the features can be distinguished more easily. In addition, we visualize the result of our spatial attention mechanism and use the middle image of each clip as the background. As shown in Fig. 5, the brightness in each image represents the attentional degree on different regions. It can be seen that our spatial attentional mechanism effectively and dynamically focuses on the most discriminative areas of the video, i.e., usually the bottle and movement of the hand.

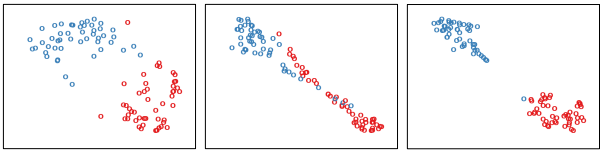

Fig. 4. Visualization of features captured by C3D, B3D and SA-B3D models using t-SNE [21] method.

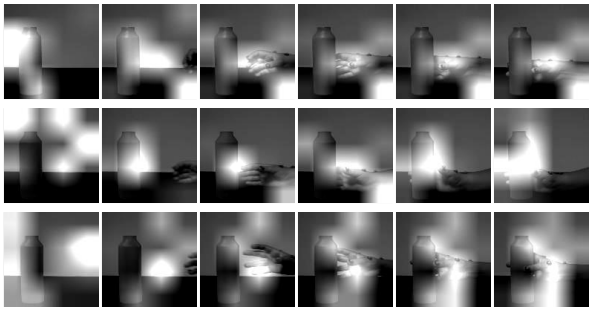

Fig. 5. Visualization of spatial attention results. Six images in every line represent the attentional results of six clips in each sample video.

5. CONCLUSION

In this paper, we propose a simple but effective method to enhance spatial information extraction of bilinear pooling without significantly increasing the parameters. For the detection of autism spectrum disorders, we propose a video-level model named spatial attentional bilinear 3D convolutional network with LSTM. The model outperforms our baseline with 9.88% average accuracy improvement and 7.69% detection accuracy improvement, and outperforms other state-of-the-art methods with 7.06% average accuracy improvement and 10.25% detection accuracy improvement. Without loss of generality, our proposed spatial attentional bilinear pooling can be used in models that adopt the normal bilinear pooling method.

6. ACKNOWLEDGMENT

This work is supported by the National Key Research Project of China under Grant No. 2017YFF0210903 and the National Natural Science Foundation of China under Grant Nos. 61371147 and 11433002.

7. REFERENCES

[1] Paul Scovanner, Saad Ali, and Mubarak Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 357–360.

[2] Limin Wang, Yu Qiao, and Xiaoou Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4305–4314.

[3] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[4] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[5] Joao Carreira and Andrew Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.

[6] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[7] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji, "Bilinear cnn models for fine-grained visual recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1449–1457.

[8] Tsung-Yu Lin and Subhransu Maji, "Improved bilinear pooling with cnns," *arXiv preprint arXiv:1707.06772*, 2017.

[9] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You, "Hierarchical bilinear pooling for fine-grained visual recognition," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 574–589.

[10] Andrea Zunino, Pietro Morerio, Andrea Cavallo, Caterina Ansuini, Jessica Podda, Francesca Battaglia, Edvige Veneselli, Cristina Becchio, and Vittorio Murine, "Video gesture analysis for autism spectrum disorder detection," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3421–3426.

[11] Catherine Lord, Susan Risi, Linda Lambrecht, Edwin H. Cook, Bennett L. Leventhal, Pamela C. DiLavore, Andrew Pickles, and Michael 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*, vol. 30, no. 3, pp. 205–223, Jun 2000.

[12] Shikhar Sharma, Ryan Kiros, and Ruslan Salakhutdinov, "Action recognition using visual attention," *arXiv preprint arXiv:1511.04119*, 2015.

[13] Y. Tian, X. Min, G. Zhai, and Z. Gao, "Video-based early asd detection via temporal pyramid networks," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, July 2019, pp. 272–277.

[14] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell, "Compact bilinear pooling," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[15] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You, "Hierarchical bilinear pooling for fine-grained visual recognition," *CoRR*, vol. abs/1807.09915, 2018.

[16] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[17] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville, "Describing videos by exploiting temporal structure," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4507–4515.

[18] Kan Chen, Jiang Wang, Liang-Chieh Chen, Haoyuan Gao, Wei Xu, and Ram Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[19] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[20] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[21] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.