Multimodal Deep Neural Network for Behavior Recognition with FOS-R for Children with ASD

Zhenhao Zhao[1], Eunsun Chung[2*], Kyong-Mee Chung[2], and Chung Hyuk Park[1]
[1]*Assistive Robotics and Tele-Medicine (ART-Med) Lab., Department of Biomedical Engineering*
The George Washington University, Washington, DC, 20052, USA
[2]*Digital Mental Health Lab., Department of Psychology. Yonsei University,* Seoul, Republic of Korea
[*]*Poster Presenter*

I. Introduction

The prevalence of Autism Spectrum Disorder (ASD) has been steadily increasing over the past decades, becoming a significant global concern [Chiarotti and Venerosi, 2020], [World Health Organization,], [Hertz-Picciotto and Delwiche, 2009]. Current statistics in the United States indicate that 1 in 36 children is diagnosed with ASD [Maenner et al., 2023]. Children with autism often experience symptoms such as somatosensory disturbances and atypical developmental patterns, which greatly impact daily social functioning [Mayes and Calboun, 1999].

The Family Observation Schedule (FOS) serves as a comprehensive tool for assessing family interactions across various contexts. Within the realm of autism research, FOS plays a crucial role in identifying and evaluating family interactions, offering valuable insights for the diagnosis, treatment, and support of children with autism by examining their social contexts and dynamics [Lee and Chung, 2016]. Presently, FOS data is encoded manually, a time-consuming and labor-intensive process. An automated FOS encoding algorithm could alleviate the burden on clinicians and researchers, ultimately benefiting numerous children with autism.

This study proposes an effective multi-modal deep learning model for the automated encoding of FOS data, utilizing videos of children on the autism spectrum engaged in specific tasks as training data for the model. The multi-modal data incorporated in this research includes video, audio, and task information data, among others. While previous studies have primarily overlooked the presence of task information data, our research aims to optimize model input, enabling the deep learning model to analyze task information data more effectively.

II. Method

A. Dataset description

Our dataset includes 216 videos of children with autism or their parents/guardians. There are some variants, but mostly composed 3 sets of 10 minutes. It is designed particularly for research involving children aged between 1 to 12 years of age, but can be modified for use with older children. The children in the videos performed three different tasks: (1) playing with specific types of toys, (2) performing a series of specific instructions, and (3) free playing alone. These videos are encoded every 10 seconds to describe the interaction styles (IS) of the children and their parents. These IS can be used as labels for training deep learning model.

There are 23 interaction types, with some describing the IS of parents (Praise (P) and Affection (AF)) and some other the IS of children (Non-compliance (NC) and Oppositional (OP)). And a part of IS presents a positive or negative symbol behind it describing the emotion of IS, such as SA+ for the positive social attention and SA- for the adverse social attention.

B. Data preprocessing and model training

Initially, we segmented the original video into 10-second intervals for each encoding and performed data augmentation to address any dataset imbalance. Our model training comprised of three sequential steps based on the respective modalities: vision, language, and a combined vision-language modality. The vision modality labels were primarily identified by action-based information, such as Physical Negative (PN), while the language modality was identified through language-based information, such as Praise (P). For the combined modality, both vision and language information were utilized to predict the target variable.

III. Initial result

Currently, we have accomplished the data preprocessing with their respective labels and developed an initial visual information classification prediction model, utilizing the fast slow ResNet architecture [Feichtenhofer et al., 2019], attaining an overall accuracy exceeding 70 %. Additionally, we have initiated the construction of a language model and a multi-modal model. However, the issue of model prediction bias caused by imbalanced data distribution still persists and requires further attention. As part of our future research, we aim to address this issue and enhance the overall accuracy of the model.

REFERENCES

[Chiarotti and Venerosi, 2020] Chiarotti, F. and Venerosi, A. (2020). Epidemiology of autism spectrum disorders: A review of worldwide prevalence estimates since 2014. Brain Sci, 10(5):274.

[Feichtenhofer et al., 2019] Feichtenhofer, C., Fan, H., Malik, J., and He, K. (2019). Slowfast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV).

[Hertz-Picciotto and Delwiche, 2009] Hertz-Picciotto, I. and Delwiche, L. (2009). The rise in autism and the role of age at diagnosis. Epidemiology (Cambridge, Mass.), 20(1):84.

[Lee and Chung, 2016] Lee, M. and Chung, K. (2016). Development of parent child interaction-direct observation checklist (pci-d) for children with developmental disabilities. Journal of Rehabilitation Psychology, 23(2):367–395.

[Maenner et al., 2023] Maenner, M. J., Warren, Z., Williams, A. R., and et al. (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years — autism and developmental disabilities monitoring network, 11 sites, united states, 2020. MMWR Surveill Summ, 72(No. SS-2):1–14.

[Mayes and Calboun, 1999] Mayes, S. D. and Calboun, S. L. (1999). Symptoms of autism in young children and correspondence with the dsm. Infants & Young Children, 12(2):11–23.

[World Health Organization,] World Health Organization. Autism Spectrum Disorders Fact Sheet. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders. [Accessed: March 29, 2023].