시선 정보를 활용한 Transformer 기반 아동의 상동행동 분류

김현수, 왕우진, 박은솔, 신용구*

고려대학교

{hskim7542, wwj604, espark_82, *ygshin92}@korea.ac.kr

Transformer-based Classification of Stereotypical Behavior in Children Using Gaze Information

Hyeon-Su Kim, Woo-Jin Wang, Eun-Sol Park, and Yong-Goo Shin*

Korea University

요 약

본 논문에서는 Transformer 기반 모델을 활용하여 4가지 종류 (팔 흔들기, 머리 흔들기, 제자리 회전, 장난감 놀이)의 상동행동을 분류했다. 또한, 장난감 놀이 클래스를 효과적으로 분류하기 위해 사전 학습된 모델을 활용하여 아동의 시선 주의 정보를 추출하였고, 기존 동영상의 특징 맵과 결합했다. 두 특징 맵을 학습 초기부터 융합할 때 발생할 수 있는 네트워크의 불안정성을 방지하는 그라데이션 스케줄링 학습 기법을 제안했다. 해당 방법을 통해 장난감 놀이 클래스를 추가한 두 데이터 세트 ESBD+와 SSBD+에 대해 교차 검증 결과, 각각 89.1%, 81.6%의 정확도를 기록했다.

Ⅰ. 서 론

자폐 스펙트럼 장애는 일반적으로 생후 3년 이내에 나타나는 신경 발달 장애 중 하나이다. 제한된 관심사와 반복적인 행동을 보이며 사회적 · 정서적 의사소통에서 결함을 보이는 것이 주요 특징으로 나타난다. 특히, 상동행동은 자폐 스펙트럼 장애를 진단하는 데에 있어 핵심적인 진단 지표 중 하나로, 반복적이고 제한된 방식의 행동을 보이는 특징을 가지고 있다. 상동행동에는 대표적으로 신체 흔들기, 제자리 회전, 사물 흔들기 등의 신체적 · 물리적 기반 반복 행동이 있다. 이러한 상동행동을 포함한 자폐 스펙트럼 장애를 조기에 진단하는 것은 후일 아동의 언어 및 사회적 의사소통의 발달을 포함하여 다양한 영역의 부정적 영향을 예방할 수 있기 때문에 매우 중요하다. 자폐 관련 행동의 대표적인 진단 도구로는 ADOS (Autism Diagnostic Observation Schedule)가 있으며, 자폐 스펙트럼 장애 아동에 대한 관찰 및 평가에 널리 사용된다[2]. 그러나 이를 활용하더라도, 의료진이 장시간 녹화된 영상을 관찰하거나, 환자와 오랜 시간 같이 있으면서 나타나는 행동 특징을 분석해야 하는데, 이러한 방식은 고도의 전문성이 요구되며, 진단 과정에 많은 시간이 소요된다. 따라서 의료진의 업무 부담이 증가하고, 효율적인 대응이 어려워진다.

본 논문에서는 이러한 문제점을 해결하기 위해, Transformer 기반 모델을 통해 녹화된 영상에서 아동이 어떠한 행동을 반복하는지를 분류했다. 특히, 기존 연구[6]에서 다루지 않았던 장난감 놀이 클래스를 추가하고, 사물에 대한 아동의 시선 정보를 동영상과 융합하여 관심 여부를 반영했다. 또한, 학습 초기 단계에서 혼합 데이터로 인한 백본 모델의 불안정한 학습을 방지하기 위해 그라데이션 스케줄링 학습 기법을 도입한 결과, 두 데이터 세트 간 교차 검증 정확도를 각각 89.1%와 81.6% 기록했다.

Ⅱ. 본 론

2.1. 시선 정보 기반 상동행동 분류 모델

본 연구에서는 입력된 아동의 동영상 속 상동행동의 종류를 분류하기 위하여 Video Swin Transformer[3]를 백본 네트워크로 사용하여 모델을 학습하였으며, 전체 모델의 구조는 그림 1에 제시했다. 특히 유사한 행동을 보이는 팔 흔들기 클래스와 장난감 놀이 클래스를 분류하기 위해선 아동과 사물 간 상호작용에 대한 정보를 분석할 수 있는 모듈을 추가했다.

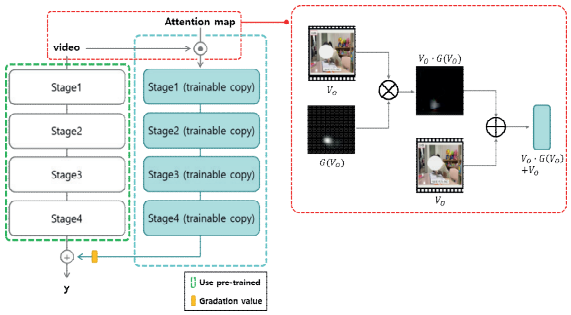

그림 1. 상동행동 검출을 위한 VST 기반 모델 아키텍처

$$f_A = (V_o \bullet G(V_o)) + V_o \quad (1)$$

식 (1)은 영상 속 아동의 머리 위치를 기반으로 시선을 예측하는 Attention target detection 모델[1]을 활용하여, 아동이 주시하는 사물에 대한 정보를 융합한 특징 맵을 정의한 수식이다. V_O는 원본 동영상, $G(\circ)$는 Attention target detection 모델을 뜻하며, $V_O \bullet G(V_O)$의 형태로 아동이 바라보는 대상을 추출한다. 이 결과는 원본 동영상 V_O에 더해져 아동이 바라보는 대상에 대해 픽셀값을 강조함으로써 시선 정보를 반영하고자 했다. 그러나, Transformer block을 통과한 시선 융합 특징 맵 f_A가 기존의 백본 네트워크를 통해 얻은 특징 맵 f_O와 직접적으로 결합될 경우, 불안정성을 유발할 수 있다. 이를 완화하기 위해, 학습 초기에는 추가 특징 맵의 반영도를 0으로 설정하고, 학습이 진행됨에 따라 해당 반영도를 점진적으로 1에 수렴시키는 그라데이션 스케줄링 학습 기법을 제안했다.

행동 / DB	팔 흔들기	머리 흔들기	제자리 회전	장난감 놀이
SSBD+	121	46	39	39
ESBD+	246	66	100	54

표 1. 상동행동 분류를 위한 데이터 세트 구조

$$f_I = f_O + \frac{t}{N_T} f_A \tag{2}$$

식 (2)는 그라데이션 스케줄링 학습 기법을 설명하는 수식으로, 앞서 언급한 시선 융합 특징 맵 f_A가 기존 백본 네트워크에서 추출된 특징 맵 f_O와 결합되어 학습 시 불안정성을 방지하기 방법을 제시한다. 여기서 N_T는 전체 학습 반복 횟수, t는 현재 학습 횟수를 나타낸다. 이 기법을 통해, 추가된 시선 융합 특징 맵의 영향을 점차 증가시키며, 학습 초기의 불안정한 학습을 방지할 수 있다.

2.2. 실험 환경

본 연구에서는 자폐 스펙트럼 장애 아동의 상동행동을 분류하기 위해 SSBD (Self-Stimulatory Behaviour Dataset)[5]와 ESBD (Expanded Stereotyped Behaviour Dataset)[4]를 기반으로 학습 데이터를 구축했다. SSBD는 3개 클래스 (팔 흔들기, 머리 흔들기, 제자리 회전), ESBD는 4개 클래스 (팔 흔들기, 머리 흔들기, 제자리 회전, 손동작)로 구성되어 있으며, 각 동영상의 상동행동이 관찰되는 부분을 추출하여 206개, 412개의 샘플로 분할하였다. 이에 사물을 이용한 상동행동 영상을 YouTube 크롤링을 통해 33개를 수집하여 장난감 놀이 클래스 93개의 샘플을 추가하였다. 추가된 샘플을 SSBD와 ESBD에 추가하여 이를 SSBD+ (245개), ESBD+ (466개)로 통합하여 표 1과 같이 확장된 데이터 세트를 구축하였다.

본 연구에서 제안한 시선 기반 상동행동 분류 알고리즘의 성능을 평가하기 위해, SSBD+와 ESBD+ 데이터 세트를 활용하여 교차 검증을 수행했다. 백본 네트워크로는 앞서 언급하였듯이 Video Swin Transformer를 사용하였고, 옵티마이저는 SGD, 초기 학습률 (Initial learning rate)은 1×10^{-4}, 가중치 감쇠(Weight decay)는 1×10^{-4}로 설정하여 총 50 epoch 동안 학습을 진행했다. 실험은 NVIDIA RTX 4090 GPU 환경에서 수행되었으며, 모델의 성능은 정확도와 F1 score를 기준으로 평가했다.

2.3. 실험 결과

제안한 알고리즘의 성능을 정량적으로 검증하기 위하여 SSBD+와 ESBD+에 대해 교차 검증을 수행하였다. 또한, 시선 정보 융합 및 그라데이션 스케줄링 학습 기법의 유무에 따른 성능 차이를 표 2에 제시하였으며, 각 기법의 적용 여부에 따라 모델의 분류 성능이 어떻게 달라지는지를 비교 · 분석하였다. 그 결과, 두 기법을 모두 적용했을 때 SSBD+와 ESBD+에서 검증 시 가장 높은 80.80%, 88.80%의 정확도와 77.25%, 87.43%의 F1 score를 달성했다. 이를 통해 시선 정보를 반영한 것이 아동과 사물 간 상호작용을 효과적으로 포착하였을 뿐만 아니라, 점진적인 반영을 통해 모델이 아동의 행동과 주위 환경에 대한 맥락 정보를 반영하여 기존 방식보다 더욱 정밀한 분류가 가능해졌음을 확인할 수 있다.

Ⅲ. 결 론

본 논문에서는 자폐 스펙트럼 장애의 핵심 진단 지표인 상동행동을 자동으로 분류하기 위해, 아동의 시선 정보를 활용하는 알고리즘을 제안했다.

학습/검증	시선 융합	그라데이션 스케줄링 학습	정밀도 (%)	재현율 (%)	F1 score (%)	정확도 (%)
ESBD+/ SSBD+			84.93	71.98	73.33	80.00
	✓		76.70	69.35	71.78	76.30
		✓	80.75	75.70	**77.25**	**80.80**
SSBD+/ ESBD+			86.38	82.25	84.03	86.20
	✓		80.38	82.70	80.75	87.40
		✓	87.28	89.3	**87.43**	**88.80**

표 2. 시선 융합 및 그라데이션 스케줄링 학습 기법에 따른 학습 결과

또한, 기존 ESBD와 SSBD에 사물 기반 상동행동 데이터를 추가하여 ESBD+와 SSBD+를 구축하였으며, 시선 정보와 원본 동영상을 효과적으로 결합하기 위하여 그라데이션 스케줄링 학습 기법을 도입했을 때 가장 우수한 성능을 기록했다. 본 연구의 한계점으로는 의료 데이터의 특성상 확보 가능한 학습 데이터의 규모가 한정적이라는 점과 원본 동영상과 시선 정보를 독립적으로 분석하는 구조로 인해 연산량이 두 배 이상 증가한다는 점이 있다. 향후 연구에서는 두 가지 정보를 동시에 처리하는 네트워크를 도입하여 연산량을 줄임으로써 실제 임상에서의 활용 가능성이 증가할 것으로 기대한다.

ACKNOWLEDGMENT

This work was supported by IITP grant funded by MSIT (Grant#. RS-2025-02263277, 50%). This work was supported by ITRC support program supervised by the IITP and funded by MSIT (Grant#: IITP-2025-RS-2023-00258971, 50%)

참 고 문 헌

[1] E. Chong, Y. Wang, N. Ruiz, and J.M. Rehg. “Detecting attended visual targets in video.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[2] C. Lord, M. Rutter, S. Goode, H. Heemsbergen, H. Jordan, L. Mawhood, and E. Schopler. "Austism diagnostic observation schedule: A standardized observation of communicative and social behavior." Journal of autism and developmental disorders 19.2 (1989): 185-212.

[3] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu. "Video swin transformer." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

[4] F. Negin, B. Ozyer, S Agahian, S. Kacdioglu, and G.T. Ozyer. “Vision-assisted recognition of stereotype behaviors for early diagnosis of autism spectrum disorders.” Neurocomputing 446 (2021): 145-155.

[5] S. Rajagopalan, A. Dhall, and R. Goecke. "Self-stimulatory behaviours in the wild for autism diagnosis." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013.

[6] C.H. Yoo, J.H. Yoo, M.K. Back, W.J. Wang, and Y.G. Shin. "A unified framework to stereotyped behavior detection for screening Autism Spectrum Disorder." Pattern Recognition Letters 186 (2024): 156-163.