시계열 랜드마크 전처리와 입력 특징 확장을 이용한 Multi Encoder 구조 한국어 수어 단어 인식

장하민, 박현희*
명지대학교 정보통신공학과

gkals4054, hhpark*@mju.ac.kr

Korean Sign-Language(KSL) Word Recognition Using a Multi Encoder Architecture with Temporal Landmark Preprocessing and Input Feature Expansion

Hamin Jang, Hyunhee Park*
Dept. Information and Communication Engineering, Myongji University

요 약

본 논문에서는 AI Hub 에서 제공하는 대규모 한국어 수어 단어 데이터셋을 활용하여 multi encoder 기반 수어 단어 인식 모델을 학습하고, 시계열 랜드마크 입력에 대한 전처리 및 특징 확장을 통한 성능 향상을 분석한다. 좌표 정규화, 스케일링, 속도 및 가속도 특성 추가와 안정적인 하이퍼파라미터 설정을 통해 기존 선행연구 대비 향상된 단어 인식 성능을 보였으며, 최대 96.2%의 정확도를 기록하였다.

Ⅰ. 서 론

수어(sign language)는 청각장애인(농인)이 일상 생활에서 타인과 의사소통하기 위해 사용하는 필수적인 언어 수단으로, 농인 사회의 핵심적인 의사소통 도구이다. 그러나 수어를 이해할 수 있는 비농인의 비율은 매우 낮으며, 전문 통역 인력의 공급 부족과 높은 비용, 시간·공간적 제약 등의 이유로 청각장애인은 일상생활, 공공 서비스 및 응급 상황에서 지속적인 의사소통 장벽을 겪고 있다. 이러한 문제를 완화하기 위하여 컴퓨터 비전과 딥러닝 기술을 활용한 수어 인식 연구가 꾸준히 진행되어 왔다.

그러나 상당수의 기존 수어 인식 연구들은 지문자 위주의 단순 정지 이미지 분류에 그치거나, 연구자가 직접 촬영한 소규모의 수어 단어 영상을 중심으로 학습을 진행해왔다[1][2]. 이러한 접근은 의사소통에 필요한 수어의 다양하고 연속적인 동작 패턴과 시간적 흐름을 충분히 반영하지 못한다는 한계를 갖는다.

이에 본 연구에서는 AI Hub 에서 제공하는 대규모 한국어 수어 단어 데이터셋을 활용하여 기존 연구에서 제안된 multi encoder 수어 인식 구조를 기반으로 시계열 랜드마크 데이터의 표현력과 학습 안정성을 향상시키기 위한 다양한 전처리 및 학습 기법을 적용한다[3]. 프레임 간 좌표 변화량을 이용한 속도 및 가속도 특성 추가, 영상 전체 프레임의 손목 간 거리 평균을 기준으로 한 스케일 정규화 등의 기법을 적용하여 기존 연구 대비 향상된 단어 인식 성능을 확인하였다.

Ⅱ. 본론

본 연구에서는 AI Hub에서 제공하는 대규모 한국어 수어 데이터셋을 활용하여 수어 단어 인식 모델을 학습한다. 해당 데이터셋은 단어, 문장, 지문자 데이터를 포함하며 다수의 수어 제공자와 촬영 각도에서 수집된 대규모 영상과 해당 영상 30fps 분할 이미지에 대한 랜드마크(관절정보) 좌표 값으로 구성되어 있다. 이 중 단어 3,000개에 대하여 15명의 수어 제공자에게서 촬영된 총 45,000개 정면 각도 단어에 대한 좌·우 손 랜드마크 좌표 데이터를 활용한다. 각 프레임은 왼손과 오른손 각각 21개의 랜드마크로 구성된 총 42개의 랜드마크 좌표를 포함하고, 각 랜드마크 값은 픽셀 좌표계 기준의 (x, y)값과 검출 신뢰도를 의미하는 confidence 값으로 이루어져 있다. 학습 과정에서는 confidence 값을 제외한 좌표 정보만을 사용한다.

그림 1. 수어 영상 데이터 프레임 예시

모델 입력은 프레임 단위 시퀀스로 구성되며 각 프레임의 42개 관절 정보를 펼쳐 하나의 벡터 토큰 입력으로 사용한다. 이를 통해 전체 입력 시퀀스는 시간 축을 따라 나열된 토큰들의 집합으로 표현되며 transformer 구조를 통해 시계열 패턴을 학습하도록 설계된다. 입력 벡터는 선형 변환을 통해 모델의 임베딩 차원으로 투영된 뒤 sin/cos 함수 기반의 positional encoding 이 더해져 시간 정보를 반영한다.

본 연구에서 사용한 모델 구조는 기존 선행연구에서 제안된 multi encoder 기반 transformer 구조를 기본 모델로 한다[4]. 해당 구조는 전체 관절 정보를 입력으로 사용하는 global branch 와 왼손 및 오른손 관절을 각각 입력으로 사용하는 두 개의 single hand branch 로 구성된다. 각 branch 는 동일한 구조의 transformer encoder 를 사용하며 encoder 는 다층 self-attention 과 feed-

forward network 로 구성된다. 본 연구에서는 encoder layer 수를 4로 설정하고 global branch 에는 8개의 attention head 를, 좌·우 single hand branch 에는 각각 4개의 head 를 사용한다. 각 branch 의 출력은 시간 축에 대해 average pooling 을 수행한 뒤 layer normalization 을 거쳐 분류 헤드로 전달되며 최종 예측은 global branch, 왼손·오른손 single branch 의 logit 값을 합산하는 방식으로 결정된다.

기본 모델에서는 프레임 단위의 좌표 정보만을 입력으로 사용하였으나, 본 연구에서는 시계열 표현력을 강화하고 일반화 성능을 향상시키기 위해 다양한 전처리 및 입력 확장 기법을 단계적으로 적용한다. 먼저 좌표 분포의 차이를 줄이기 위해 각 프레임마다 왼손과 오른손 손목 관절의 중점을 기준으로 모든 관절 좌표를 평행 이동한다. 이후 시퀀스 전체에서 계산한 손목 간 평균 거리를 기준으로 스케일 정규화를 수행함으로써 촬영 거리 및 손 크기 차이에 대한 영향을 완화시킨다. 또한 기존 좌표 정보(p_t)에 동적 특성을 추가로 반영하기 위하여 프레임 간 좌표 차이를 이용한 속도 특성(v_t)과 속도 변화량을 이용한 가속도 특성(a_t)을 입력 특징으로 추가한다.

$$v_t = p_t - p_{t-1}, \qquad a_t = v_t - v_{t-1} \tag{1}$$

이를 통해 각 관절은 (x, y) 좌표뿐 아니라 이동 방향과 변화 패턴을 함께 포함하는 6차원 특징 벡터로 확장되었으며, 모델 입력 차원은 관절당 6차원으로 구성된다.

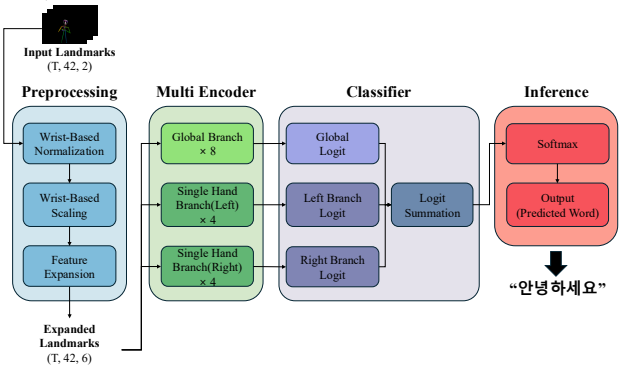

그림 2. Multi Encoder 모델 구성도

모델 학습에는 cross-entropy loss 를 사용하며, 옵티마이저로는 AdamW 를 적용한다. 또한 OneCycleLR 스케줄러를 통해 학습 초반의 빠른 수렴과 안정적인 후반 학습을 유도한다. 과적합 방지를 위해 dropout 과 early stopping 기법을 적용하고, 대규모 데이터셋을 효율적으로 학습하기 위해 사전에 생성한 catalog 파일을 기반으로 배치 단위 데이터 로딩 방식을 사용한다. 총 학습 epochs 는 100, batch size 는 64, learning rate 는 1e-3 으로 설정했으며 early stopping 의 patience 는 7로 설정한다.

Ⅲ. 실험 결과

실험 결과, 제안한 전처리 및 입력 확장 기법을 모두 적용한 모델은 96.2%의 검증 데이터 정확도로 기존 선행연구에서 나타낸 91.0%의 결과를 상회하는 결과를 기록하였다. 이는 전처리된 프레임 정보와 속도 및 가속도 특성이 모델 학습에 효과적으로 활용되었음을 보여준다. 반면 실제 환경에서 촬영된 수어 영상에 대해서는 성능 저하가 관찰되었으며, 이는 모델 구조 자체의 한계보다는 학습 데이터와 실제 촬영 데이터 간의 분포 차이, 랜드마크 품질 저하, 동작 속도 변화 및 잡음 프레임의 영향에 기인한 것으로 분석된다.

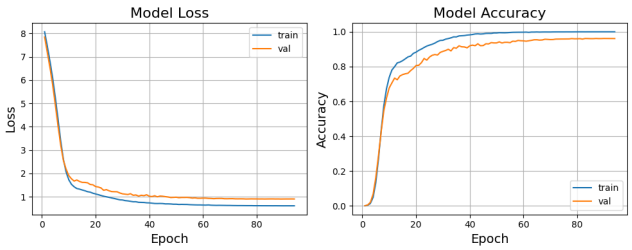

그림 3. ME + norm/scale + vel/acc 모델 학습 결과

표 1. 각 전처리 기법에 따른 모델 학습 결과 비교

Method	Val acc	Val loss
Single encoder	0.776	-
Baseline	0.910	-
Multi encoder (ME)	0.946	0.983
ME + norm/scale	0.960	0.911
ME + vel/acc	0.954	**0.884**
ME + norm/scale + vel/acc	**0.962**	0.901

Ⅳ. 결론

본 연구에서는 AI Hub 에서 제공하는 대규모 한국어 수어 단어 데이터셋을 활용하여 multi encoder 기반 수어 인식 모델을 학습하였다. 시계열 랜드마크 입력에 대해 좌표 정규화 및 스케일링, 속도 및 가속도 특성 추가 등 전처리 및 입력 확장 기법을 적용하고, 학습률, 배치크기, 학습 스케줄러 등 주요 하이퍼파라미터를 안정적으로 설정함으로써 단어 인식 성능의 향상을 확인하였다.

향후 연구에서는 학습 데이터와 실제 입력 데이터 간의 분포 차이를 완화하고, 다양한 촬영 환경과 동작 변화를 반영한 학습 전략을 적용함으로써 실제 촬영 환경에서도 신뢰성 있게 동작하는 수어 인식 시스템의 구현을 목표로 한다.

ACKNOWLEDGMENT

본 과제(결과물)는 2025 년도 교육부 및 경기도의 재원으로 경기 RISE 센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-09-A15)

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "수어 영상"을 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

참 고 문 헌

[1] 공준석, 김정이, "Mediapipe Holistic 을 활용한 TFLite 기반 수어 인식 모델 개발," 문화기술의 융합, vol. 11, no. 5, pp. 271-278, 2025.

[2] 이계진, 표승현, 김승현, 김건우, 오준수, 김혜윤, "MLP 모델과 LSTM 모델을 활용한 수어 번역기 구현," 대한전기학회 학술대회 논문집, pp. 2814-2815, 2025.

[3] AI Hub, 수어 영상, (https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=data&dataSetSn=103)

[4] 노경근, 황의준, 이희제, 박종철, "트랜스포머 모델을 사용한 특징점 학습과 한국어 수어 인식," 한국정보과학회 학술발표논문집, pp. 1074-1076, 2022.