멀티모달 기반 **119** 신고 상황 인식 및 다중 정보 자동 추출 시스템 최종 보고서

박지혜, 박주연, 김윤서 (팀명: 삐용삐용)

20221367, 20221365, 20221337

“Multimodal 119 Emergency Call Understanding and Information Extraction System”

Park Ji Hye, Park Ju Yeon, Kim Yun Seo (Team: beep beep)
Sungshin Women’s Univ.

요 약

본 연구는 실제 119 신고 음성 데이터를 바탕으로, 재난 상황의 유형(총 4가지)과 긴급도(3단계)를 동시에 분류할 수 있는 멀티모달 기반의 다중 태스크 딥러닝 모델을 개발하는 데 목적을 두고 있다. 기존의 119 신고 접수 시스템은 대부분 접수자의 주관적인 판단에 의존하고 있으며, 이로 인해 발생할 수 있는 판단 오류를 줄이기 위한 자동화 모델의 필요성이 제기되어 왔다. 이를 해결하기 위해 본 연구에서는 음성 스크립트의 텍스트 정보와 음성 신호의 특성을 결합한 상황 인식 모델을 설계하였다. 초기에는 KoBERT 기반의 텍스트 인코더와 10개의 음성 특성을 활용한 DNN 구조를 사용하였으나, 이 방식은 과도한 클래스 수(16개)로 인한 정보 분산, 음성의 시간 흐름 미반영, 정보 손실 등의 문제점이 나타났다. 이러한 한계를 극복하고자 최종 모델에서는 입력 텍스트의 길이를 확장하고 상황 유형 라벨을 4개의 주요 분류로 단순화하였다. 텍스트는 Kc-ELECTRA 인코더를 통해 보다 깊이 있는 문맥 임베딩을 생성하였으며, 음성은 MFCC(Mel-Frequency Cepstral Coefficients) 시퀀스로 변환한 뒤 Bi-LSTM 구조를 통해 음성의 시간적 흐름을 학습할 수 있도록 설계하였다. 최종적으로는 이 두 가지 모달리티의 임베딩을 통합하고, 이를 기반으로 두 분류 태스크(상황 유형과 긴급도)를 동시에 수행하는 심층 신경망을 구성하였다. 과적합 방지를 위해 Early Stopping 기법도 도입하였으며, 이를 통해 모델의 일반화 성능을 확보할 수 있었다. 이 연구는 실제 신고 데이터를 활용하여 멀티모달 학습 모델의 실효성을 검증하였다는 점에서 의의가 있으며, 향후 119 신고 접수 시스템에서 신속하고 일관된 판단을 지원하는 도구로서 활용 가능성을 제시한다.

Ⅰ. 서 론

기존 119 신고 접수 시스템은 신고를 받는 접수자의 경험과 직관에 의존하여 신고 상황을 판단하고 긴급도를 분류하는 방식으로 운영된다. 특히, 복합적이거나 위급도가 높은 재난 상황에서는 신속하고 일관된 판단에 어려움이 발생할 수 있으며 이는 골든타임 확보에 중요한 제약이 된다. 이러한 주관적 판단의 한계를 극복하고 재난 대응의 효율성을 극대화하기 위해 신고 음성을 활용한 자동화된 상황 인식 시스템 구축의 필요성이 대두되고 있다.

본 연구는 실제 119 신고 음성 데이터를 활용하여 신고 내용의 상황 유형(4개의 클래스)과 긴급도(3개의 클래스)를 동시에 예측하는 멀티모달 심층 학습 모델을 제안한다. 이는 음성 스크립트의 텍스트 정보와 음성의

비언어적 특징(acoustic features)을 통합하여 응급 상황 대응 시간을 단축 시키고 빠르고 효율적으로 여러 신고를 처리 할 수 있게 돕는다

Ⅱ. 본론

2.1 선행연구

AI Hub에서는 실제 119 신고 데이터를 활용한 다양한 AI 모델을 제시하고 있다. 그중 텍스트 기반 신고 분류 모델은 119 신고 발화 텍스트를 입력으로 하여 Kc-ELECTRA 기반 Sequence Classification 방식으로 신고 유형을 분류한다. 해당 모델은 텍스트 정보만을 활용하여 비교적 높은 분류 성능을 보였으나, 음성 신호에 포함된 억양, 강세 등 비언어적 정보는 반영하지 못한다는 한계를 가진다.[1]

또한, 119 신고 접수자의 의사결정을 지원하기 위한 딥러닝 기반 재난 상황 인지 및 대응 지원 모델이 제안된 바 있다. 해당 연구는 119 상황관리 표준매뉴얼을 기반으로 신고 접수 초기 단계에서 재난 상황을 자동 인식하고, 접수자의 판단을 보조하는 의사결정 지원 모델을 설계하였다. 이를 통해 신고 접수자의 개인 역량에 의존하던 기존 업무 방식의 한계를 완화하고자 하였으며, 실험을 통해 제안한 모델의 유효성을 검증하였다.[2]

이러한 선행연구들은 119 신고 접수 업무의 자동화 및 의사결정 지원 가능성을 보여주었으나, 음성과 텍스트를 동시에 활용하고 시간적 특성을 반영한 통합적 접근에는 한계가 있었다. 이에 본 연구는 멀티모달 입력과 멀티태스크 학습을 결합한 모델을 통해 기존 연구의 한계를 보완하고자 한다.

2.2 데이터 구성 및 분석

본 연구에서는 AI HUB에서 제공하는 '위급상황 음성, 음향(고도화) - 119 지능형 신고접수 음성 인식 데이터'를 사용하였다. 해당 데이터는 실제 119 긴급 신고 통화를 기반으로 수집된 약 3,000시간 분량의 음성 데이터와 약 16만건의 신고 사례로 이루어져 있다. 각 신고 데이터는 WAV 형식의 음성 파일과 JSON 형식의 라벨링 데이터로 제공되며 라벨링 데이터에는 발화 구간별 전사 텍스트, 화자 정보, 통화 시작, 종료 시점 등이 포함된다. 또한 신고 상황 대분류 및 중분류, 긴급도, 감정 상태 등의 데이터가 함께 제공된다. 본 연구에서는 이 중 상황분류와 긴급도 예측을 주요 목표로 설정하였다.

데이터 분포를 대분류 기준으로 분석한 결과, 전체 158,973건의 신고 데이터 중 구급 상황이 약 75%를 차지하여 가장 높은 비중을 보였다. 반면 구조 상황은 약 14%, 화재 상황은 약 7%, 기타 상황은 약 3% 수준으로 상대적으로 적은 비중을 차지하는 것으로 확인되었다. 이러한 분포는 실제 119 신고가 발생하는 비율을 반영하는 동시에 학습 측면에서는 특정 클래스에 대한 편향되는 문제점이 있다. 이와 같은 클래스 불균형은 편향된 예측을 수행할 가능성을 높이며, 소수 클래스에 대한 분류 성능 저하로 이어질 수 있으므로 데이터 전처리를 시행하였다.

2.3 데이터 전처리

본 연구에서는 음성 및 텍스트 데이터를 효과적으로 활용하기 위해 단계적인 전처리 과정을 수행하였다. 전처리는 데이터 구조 정렬, 음성 특징 추출, 텍스트 정제 및 임베딩 생성을 중심으로 구성된다.

먼저 데이터 구조를 통일하기 위해 음성 파일과 전사 텍스트를 신고 ID 기준으로 정렬하고, 상황 라벨에 따라 폴더 구조를 재구성하였다. 이를 통해 음성 데이터와 텍스트 데이터가 동일한 샘플 단위로 관리될 수 있도록 하였으며 학습, 검증, 테스트 데이터 분할 시에도 데이터 누수를 방지하였다. 또한 4가지 대분류 기준 데이터 분포를 분석하여 클래스 간 불균형이 존재함을 확인하였으며, 이는 이후 실험 결과 해석 시 고려 요소로 활용하였다. 음성 데이터의 경우, 초기 실험에서는 에너지, 피치, 음질과 관련된 약 40개의 통계 기반 음성 특징을 추출하여 사용하였다. 그러나 해당 방식은 음성 신호의 시간적 변화를 충분히 반영하지 못해 모델 성능 향상에 한계가 있음을 확인하였다. 이에 따라 본 연구에서는 MFCC(Mel-Frequency Cepstral Coefficient)를 기반으로 한 시계열 음성 표현 방식을 채택하였다. 각 음성 신호는 프레임 단위로 분할한 후, 프레임마다 40차원의 MFCC 계수를 추출하였으며, MFCC 계수는 차원별 평균과 표준편차를 이용해 정규화하였다. 이후 음성 길이에 따른 시계열 차이를 보정하기 위해 시간축 보간을 적용하여, 모든 음성 샘플을 300 타임스텝 × 40차원의 MFCC 시퀀스 형태로 통일하였다. 이러한 시계열 음성 특징은 이후 Bi-LSTM 모델을 통해 시간적 패턴을 효과적으로 학습하도록 하였다. 텍스트 데이터에 대해서는 전사 텍스트에서 의미 없는 반복 표현과 관용적인 응답을 제거하기 위해 불용어 사전을 정의하였다. "네", "예", "어", "119입니다"와 같이 신고 접수 과정에서 빈번히 등장하지만 상황 판단에는 기여하지 않는 표현을 제거함으로써 텍스트 임베딩의 정보 밀도를 향상시켰다. 또한 상황 라벨과 긴급도 라벨을 각각 정수 형태로 변환하여 모델 학습에 적합한 형태로 전처리하였다.

최종적으로 전처리된 음성 특징과 텍스트 임베딩은 동일한 샘플 단위로 결합되어 멀티모달 입력으로 구성되었으며, 이를 통해 음성과 텍스트 정보를 동시에 고려하는 심층학습 모델 학습이 가능하도록 하였다.

2.4 모델 학습

본 연구에서는 멀티모달 다중 태스크 모델의 효과적인 학습을 위해 PyTorch 프레임워크 기반의 학습 파이프라인을 설계하였다. 초기 실험 단계에서 나타났던 모델 수렴 불안정성 문제를 해결하고, 긴급도 분류의 상대적 중요도를 반영하기 위한 손실 함수 가중치 조정 전략이 포함되었다.

2.4.1 옵티마이저 및 학습률

모델의 최적화를 위해 AdamW 옵티마이저를 채택하였다. 이 옵티마이저는 weight decay를 보다 엄밀하게 적용하여, 과적합을 방지하고 모델의 일반화 성능을 안정적으로 끌어올리는 데 효과적이다. 학습률은 초기 실험에서 모델이 불안정하게 학습되는 경향을 고려하여 기본값 $3e^{-5}$로 설정하였다.

2.4.2 손실 함수 및 가중치

상황(4 클래스)과 긴급도(3 클래스) 두 가지 분류 태스크를 동시에 학습하는 다중 출력 구조이므로, 두 개의 독립적인 손실 함수를 사용하고 긴급도 태스크의 중요도를 반영한 가중치 전략을 적용하였다.

- 손실 함수: 두 태스크 모두 다중 분류 문제이므로 교차 엔트로피 손실(CrossEntropyLoss)을 사용하였다.
- 태스크 가중치 (Task Weighting): 전체 손실 계산 시, 신고 시스템에서 중요도가 높고 신속한 판단이 요구되는 긴급도 손실에 사전 실험을 통해 가장 좋은 결과인 2.5배의 높은 가중치를 부여하였다. 이러한 가중치 설정은 모델이 긴급도 예측을 상황 예측보다 우선순위가 높은 태스크로 인식하고 학습에 집중하도록 유도한다.

2.4.3 학습 전략 및 조기 종료

모델의 최종 성능을 확보하고 일반화 능력을 평가하기 위해 두 가지 핵심 학습 전략을 비교하였다.

1. 베이스라인 (Electra Freeze): 텍스트 인코더인 Kc-ELECTRA의 가중치는 고정하고 새로 추가된 퓨전 레이어 및 분류 헤드만 학습시켰다. 이 방식은 학습 속도는 빠르지만 텍스트 인코더의 특징 표현력이 신고 데이터에 맞게 정교하게 미세 조정되지 못하는 한계를 갖는다.
2. 파인튜닝 (Electra Trainable): Kc-ELECTRA의 가중치까지 포함하여 모델 전체를 함께 학습시켰다. 이 방식은 계산 비용은 높지만, 사전 학습된 언어 모델을 현재 신고 데이터에 최적화하여 성능을 극대화할 수 있다.

모델이 훈련 데이터에 과도하게 적응하는 과적합을 방지하기 위해 Early Stopping 기법을 도입하였다. 조기 종료 기준은 검증 데이터에서의 긴급도와 정확도 향상 여부로 설정하였으며 일정 횟수(2 epoch) 동안 성능 개선이 없을 경우 학습을 중단하고, 그 시점까지 저장된 최적 가중치를 최종 모델로 사용하였다.

이와 같은 세심한 학습 전략과 안정화 기법의 도입 덕분에 훈련 손실은 지속적으로 감소하고 검증 손실은 일정 시점 이후 증가하는 경향을 보여 과적합 방지 기법이 효과적으로 작동했음을 확인했다. 결과적으로 본 모델은 실제 환경에 적용 가능한 높은 일반화 성능을 확보할 수 있었다.

2.5 실험 결과

본 연구에서는 모델 성능을 정량적으로 평가하기 위해 Accuracy, Precision, Recall, F1-score (macro / weighted)를 사용하였다.

- Accuracy는 전체 샘플 중 올바르게 분류된 비율을 의미한다.
- Precision은 모델이 특정 클래스로 예측한 결과 중 실제로 해당 클래스에 속하는 비율로, 오탐에 대한 민감도를 나타낸다.
- Recall은 실제 해당 클래스에 속하는 샘플 중 모델이 올바르게 예측한 비율로, 미탐에 대한 민감도를 의미한다.
- F1-score는 Precision과 Recall의 조화 평균으로, 두 지표 간 균형을 평가한다.
- F1-macro는 각 클래스의 F1-score를 동일한 가중치로 평균하여 클래스 불균형에 대한 영향을 확인하는 데 사용되며,
- F1-weighted는 각 클래스의 샘플 수를 고려하여 가중 평균한 지표로, 전체 성능을 대표하는 지표로 활용된다.

Metric	KC-ELECTRA (Freeze)	KC-ELECTRA (Fine-tuning)
Accuracy	0.6472	**0.8974**
F1-macro	0.1983	**0.7750**
F1-weighted	0.5101	**0.8931**
Precision	0.2334	**0.7955**
Recall	0.2506	**0.7629**

표1.상황 분류 (Major) 예측 결과 성능 비교

Metric	KC-ELECTRA (Freeze)	KC-ELECTRA (Fine-tuning)
Accuracy	0.4072	**0.5946**
F1-macro	0.3593	**0.5967**
F1-weighted	0.3650	**0.6018**
Precision	0.4106	**0.6127**
Recall	0.3966	**0.5955**

표2.긴급도 분류 (Urgency) 예측 결과 성능 비교

표 1은 Kc-ELECTRA를 고정된 임베딩으로 사용한 베이스라인 모델과 파인튜닝을 적용한 모델의 성능을 비교한 결과를 나타낸다. 상황 분류에서는 파인튜닝 적용 후 Accuracy, Precision, Recall 및 F1-score 전반에서 큰 폭의 성능 향상이 확인되었으며, 특히 F1-macro의 상승은 클래스 불균형 환경에서도 분류 성능이 안정적으로 개선되었음을 의미한다.

표 2에서는 긴급도 분류의 경우에도 파인튜닝을 통해 모든 평가지표에서 일관된 성능 향상이 나타났으나, 상황 분류에 비해 상대적으로 낮은 성능을 보였다. 이는 긴급도 라벨이 발화자의 표현 방식과 신고 맥락에 따라 판단 기준이 모호한 특성을 가지기 때문으로 해석된다. 그럼에도 불구하고, 제안한 모델은 기존 베이스라인 대비 개선된 성능을 보여주어 멀티모달 및 파인

튜닝 전략의 효과를 확인할 수 있었다.

2.6 시연 방법

제안한 멀티모달 모델의 실제 활용 가능성을 확인하기 위해, 본 연구에서는 119 신고 음성을 입력으로 하는 엔드투엔드 데모 시스템을 구현하였다. 입력된 신고 음성은 Whisper-large-v3와 학습된 모델을 사용해 상황 유형과 긴급도를 동시에 예측한다.

또한 모델의 예측 결과를 신고 접수자가 직관적으로 활용할 수 있도록 대규모 언어 모델을 활용한 후처리 모듈을 구성하였다. STT 결과로부터 위치 정보를 추출하고, 예측된 상황 분류 및 긴급도 정보를 반영하여 신고 내용을 요약함으로써 신고접수자의 의사결정을 돕는다. 데모 시스템은 Gradio 기반 웹 인터페이스로 구현되었으며, 음성 입력부터 분류 및 요약 결과까지의 전 과정을 확인할 수 있도록 설계되었다. 이러한 정성적 실험 결과를 통해 제안한 모델의 실제 적용 가능성을 확인하였다.

사진1. 시연 화면

Ⅲ. 결론

본 연구에서는 실제 119 신고 음성 데이터를 활용하여 상황 유형과 긴급도를 동시에 예측하는 멀티모달 다중 태스크 심층 학습 모델을 제안하였다. 텍스트와 음성 정보를 통합한 모델은 베이스라인 대비 전반적인 분류 성능 향상을 보였으며, 특히 Kc-ELECTRA 파인튜닝과 MFCC 기반 시계열 음성 표현의 효과를 확인하였다. 본 연구는 실제 신고 데이터를 기반으로 멀티모달 접근의 가능성을 입증하였으며, 향후 119 신고 접수 시스템의 보조 판단 도구로 활용될 수 있는 기반을 제시한다.

참 고 문 헌

[1] AIHub, 위급상황 음성/음향(고도화) – 119 지능형 신고접수 음성 인식 데이터, https://www.aihub.or.kr/aihubdata/data/view.do?currMenu115&topMenu=100&aihubDataSe=data&dataSetSn=71768

[2] Kwon E., Lee M., Park H., and Lee K.-C., "Deep Learning-based Models for Disaster Situation Awareness and Response Support," Journal of KIISE, vol. 50, no. 8, pp. 712-719, Aug. 2023.