음성 기반 상황 분류 모델의 학습 효율 향상을 위한 데이터셋 정제 기법

성시열[1] 곽준형[2] 김기원[2] 서주형[2]
인하대학교 산업경영공학과[1]
인하대학교 정보통신공학과[2]
sungsiyul1005@gmail.com, kwakjoonhyung0429@gmail.com, kimgiwon0712@gmail.com, sjh990721@naver.com

Dataset Refining Techniques for Improving Training Efficiency of Voice-Based Situation Classification Models

Siyul Sung[1] Joonhyung Kwak[2] Giwon Kim[2] Joohyung Seo[2]
Department of Industrial Management Engineering, Inha University[1]
Department of Information and Communication Engineering, Inha University[2]

본 연구는 국토교통부/국토교통과학기술진흥원의 디지털 국토정보 기술개발사업 지원으로 수행되었음 (과제번호 RS-2022-00142501).

요 약

본 연구에서는 음성 인식 기반 인공지능 모델의 성능을 향상시키기 위해 감정 분류 모델을 활용한 데이터 정제 기법을 제안한다. 연구는 AI hub의 '위급상황 음성/음향' 데이터셋과 ETRI의 '한국어 멀티모달 감정 데이터셋 2020'을 결합하여 사용하였다. 딥러닝 모델 TIM Net을 감정 분류 모델로 채택하였으며, 이진 분류 모델로는 Random Forest 분류 모델을 사용한다. 준지도 학습을 통해 학습 데이터의 과적합을 방지하고, CNN 기반의 음성 상황 분류 모델을 설계하여 성능 평가를 수행한다.

연구 결과, 제안된 데이터 정제 기법을 통해 정제된 데이터셋으로 학습한 상황 분류 모델이 원본 데이터셋에 비해 더 적은 데이터를 사용하면서도 학습의 효율성을 높였다. 이를 통해 데이터 정제 기법의 효용성을 입증할 수 있었다.

1. 서 론

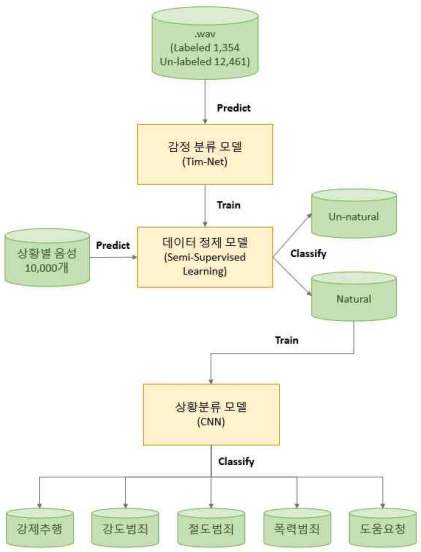
음성 인식 기반 인공지능 모델은 다양한 분야에서 놀라운 성과를 보이고 있지만, 그 성능은 대부분 학습 데이터의 질과 양에 크게 의존한다. 특히, 상황 인식 및 분류에 있어서는 고품질의 음성 데이터가 모델의 성능을 결정하는 핵심 요소로 작용한다. 그러나 현실에서는 학습 데이터가 부자연스러운 형태로 제공되는 경우가 많다. 예를 들어, AI hub의 음성 데이터를 사용하여 상황 분류 모델을 학습시키고자 했지만, 해당 데이터셋은 대부분 실제 상황에서 녹음된 것이 아니었기 때문에, 발화자의 억양과 감정에서 부자연스러운 요소가 포함되어 있었다. 이러한 요소는 모델의 성능을 크게 저하시키는 문제로 작용한다. 이를 해결하고자 감정 정보를 활용하여 자연스러운 음성을 정제하는 데이터 정제 기법을 제시한다.

2. AI hub의 "위급상황 음성/음향" 데이터셋

AI hub의 "위급상황 음성/음향" 데이터셋은 총 16가지 위급상황으로 구성되어 있다. 본 연구에서는 5가지 위급상황(강제추행, 강도범죄, 절도범죄, 폭력범죄, 도움요청)에 대한 음성 데이터를 사용하였고, 총 135,408개(188시간)로 구성되어 있다. 음성 한 개의 길이는 3초에서 5초 이내의 길이로 구성된다.

3. 본론

본 연구에서는 감정 분류 모델을 활용한 데이터 정제 기법을 사용하여 음성 기반 상황 분류 모델의 성능을 향상시킨다. 전체적인 흐름은 [그림 1]과 같다.



[그림 1] 연구 흐름도

3.1 데이터 정제 모델

본 연구에서는 감정 분류 모델의 결과를 기반으로 자연스러운 음성과 부자연스러운 음성을 구분하는 이진 분

류 모델을 구축한다.

3.1.1 데이터셋

데이터 정제 모델은 “위급상황 음성/음향” 데이터셋에서 5가지 위급상황에 대하여 각각 200개의 음성 데이터를 선별하였다. 총 1,000개의 음성에 대해 자연스러움과 부자연스러움을 0과 1로 라벨링 하였다. 이 과정에는 4명의 실험자가 참여하였으며, 주관적 판단을 최소화하기 위해 3명(75%) 이상의 동의가 있을 경우에만 라벨링을 진행하였다. 결과적으로 677개의 자연스러운 음성과 323개의 부자연스러운 음성으로 분류되었다.

추가적으로, ETRI의 ‘한국어 멀티모달 감정 데이터셋 2020’을 활용하였다. 일상 대화에 대한 음성을 데이터셋에 포함시켜 다양한 상황에서 데이터 정제 모델의 성능 향상을 도모하였다. 앞선 실험 방법을 적용하여 4명의 실험자가 부자연스러운 음성 354개를 선별하였고, 이를 통해 편향된 학습 데이터의 개수를 동등하게 맞추었다. 또한, 라벨링 하지 않은 12,461개의 음성 데이터를 확보하여 준지도 학습에 활용한다. 최종적으로 라벨링을 진행한 결과는 <표 1>과 같다.

<표 1> 데이터셋 라벨링 결과

<table>
<tr><th rowspan="2">라벨</th><th colspan="2">labeled</th><th rowspan="2">unlabeled</th></tr>
<tr><th>0 (677개)</th><th>1 (677개)</th></tr>
<tr><td rowspan="2">데이터셋</td><td>위급상황
(323개)</td><td rowspan="2">위급상황
(677개)</td><td rowspan="2">일상대화
(12,461개)</td></tr>
<tr><td>일상대화
(354개)</td></tr>
<tr><td>총 개수</td><td colspan="2">1,354개</td><td>12,461개</td></tr>
</table>

3.1.2 데이터 전처리

이진 분류 모델의 입력값으로 음성 데이터의 감정 벡터를 활용한다. 음성 데이터를 MFCC[5]를 통해 특징 벡터화를 진행하고, 음성 감정 분류 분야에서 SOTA를 달성하고 있는 TIM-Net[1]을 활용하여 감정을 분류한다. 감정 분류 모델의 결과는 데이터 정제 모델의 입력값으로 사용되기 때문에 Soft label을 사용한다. 감정 분류의 결과는 [그림 2]과 같다.

	label	file_name	angry	disgust	fear	happy	neutral	sad	surprise
0	0.0	Sess01_script01_User001F_001.wav	0.108658	0.123759	0.163147	0.077577	0.355532	0.090270	0.081057
1	0.0	Sess01_script01_User001F_002.wav	0.092113	0.130429	0.216745	0.063603	0.325484	0.094164	0.077462
2	0.0	Sess01_script01_User001F_003.wav	0.041212	0.092273	0.130359	0.033353	0.594373	0.059997	0.048433
3	0.0	Sess01_script01_User001F_005.wav	0.072157	0.108847	0.246469	0.051413	0.412226	0.056471	0.052418
4	0.0	Sess01_script01_User001F_006.wav	0.114871	0.192869	0.152400	0.091673	0.169726	0.177319	0.101141

[그림 2] 감정 분류 모델 결과

3.1.3 준지도 학습(Semi-Supervised Learning)

학습 데이터의 양을 늘리고 과적합을 방지하기 위해 준지도 학습[4] 기법을 적용하였다. 이를 위해 1,354개의 직접 라벨링된 학습 데이터와 12,461개의 라벨링되지 않은 음성 데이터의 감정 분류 결과를 활용하였다.

먼저, 지도학습 모델을 선정하기 위하여 이진 분류 모델에서 높은 성능을 보이는 3가지 분류 모델[2](Linear Regression, Random Forest, XGBoost)을 실험 대상으로 선정하였다. 과적합을 방지하기 위해 기본 파라미터를 사용하여 학습을 진행하였다. 본 연구에서는 82%의 정확도를 가진 Random Forest 분류 모델을 이진 분류 모델로 채택하였다.

위 모델을 통해 uncertainty 기반 준지도 학습을 수행하였다. 12,461개의 라벨링되지 않은 데이터를 모델을 통해 예측하였으며, 결과는 자연스러움 또는 부자연스러움의 확률로 산출되었다. 이 두 확률의 합은 1이 되며, 더 높은 값의 라벨을 예측 라벨로 선정하였다. 예측된 라벨의 확률을 예측 신뢰도(confidence)로 정의한다.

예측된 데이터 중 신뢰도 상위 5%를 신뢰성 있는 데이터로 분류하고, 이를 지도학습 데이터에 추가하였다. 확장된 데이터셋을 이용하여 지도학습을 반복하여 진행하였다. 예측 신뢰도가 낮은 2,421개(하위 20%)의 데이터는 학습에서 제외하였다. 실험 결과, 예측 신뢰도 95% 이상의 데이터 10,040개와 직접 라벨링한 데이터 1,354개를 포함하여 총 11,394개의 데이터로 데이터 정제 모델을 구축하였다.

3.1.4 정제 결과

본 연구에서 제안하는 데이터 정제 모델을 활용하여 상황 분류 모델에 사용할 데이터셋을 정제하였다. 2장에서 언급한 AI hub의 “위급상황 음성/음향“ 데이터셋에서 5가지 위급상황에 대해 각각 2,000개씩 음성 데이터를 선별하여 총 10,000개의 음성 데이터를 정제한다. 정제 과정을 거친 결과, 2,368개의 부자연스러운 음성이 제거되었으며, 정제된 결과는 <표 2>와 같다.

<표 2> 데이터 정제 결과

위급상황 라벨	정제 전 (개)	정제 후 (개)
강제추행(성범죄)	2,000	1,560
강도범죄	2,000	1,464
절도범죄	2,000	1,496
폭력범죄	2,000	1,530
도움요청	2,000	1,582
총 합계	10,000	7,632

3.1.5 정제 데이터 청각 실험

본 연구에서는 4명의 평가자가 참여하여 정제된 2,368개의 부자연스러운 음성을 검증하는 실험을 수행한다. 이 실험에서는 음성을 듣고 부자연스럽다고 느낀 인원을 동의 인원으로 정의한다. 실험 결과는 <표 3>에 나타나 있다. 동의 인원이 3명(75%) 이상인 경우의 비율이 93.5%로, 높은 정확도로 정제되었음을 확인할 수 있다. 이러한 결과는 정제 모델의 성능이 우수하다는 것을 시사한다.

<표 3> 정제 데이터셋 라벨링 실험 결과

동의 인원	4	3	2	1	0	total
음성 수	1,483	730	130	24	1	2,368

3.2 상황 분류 모델

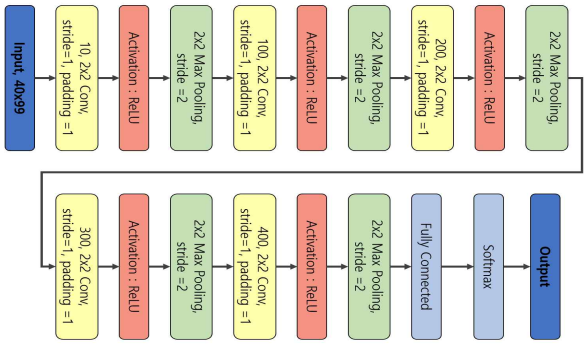
본 연구에서는 음성 상황 분류 모델을 설계하여 데이터 정제 모델의 성능을 평가한다.

3.2.1 데이터셋

<표 3>의 두 가지 데이터셋(정제 전, 정제 후)을 학습 데이터로 사용하였다. 각각 10,000개와 7,632개의 데이터를 포함하며, 8:2의 비율로 train, validation 데이터셋으로 구성하였다. 테스트 데이터셋은 "위급상황 음성/음향" 데이터셋에 포함된 테스트셋 중 각 라벨별 500개의 음성을 사용한다.

3.2.2 CNN 상황 분류

MFCC 방식을 통해 음성 데이터를 특징 벡터화하여 CNN[3] 입력으로 사용하며, 5개 층으로 이루어진 기본 CNN 모델을 구축하여 사용한다. 동일한 조건에서 데이터셋만 다르게 한 채 학습된 상황 분류 모델의 성능을 종합적으로 비교하고 데이터 정제 모델의 유의성을 평가한다. 상황 분류를 위한 CNN 모델의 구조와 기본 인자는 [그림 3]과 같다.



[그림 3] 상황 분류 CNN 모델

3.2.3 결과

상황 분류 모델의 학습 조건과 비교 결과는 <표 4>와 같다. 실험 하드웨어로 NVIDIA GeForce RTX 3080를 사용하였다.

상황 분류 모델	정제 전	정제 후
학습 데이터 수 (개)	10,000	7,632
학습 조건	epochs : 50 batch size : 10 learning rate : 0.0001 activation function : Relu optimizer : Adam	
학습 소요 시간 (초)	142.04	105.91
Accuracy	0.94	0.94
F1-Score	0.93	0.93

<표 4> 데이터 정제 평가 지표

4. 결론 및 향후 연구

본 연구는 감정 분류 모델을 활용한 데이터 정제 기법을 제안하고, 이를 음성 기반 상황 분류 모델에 적용하여 성능 평가를 수행한다. 연구 결과로 얻은 주요 발견은 다음과 같다.

제안된 데이터 정제 기법을 통해 정제된 데이터셋으로 학습한 상황 분류 모델은 더 적은 양의 데이터로 동일한 성능과 빠른 학습 속도를 보였다.

본 연구의 성과는 음성 기반 상황 분류 모델의 학습 효율 개선에 기여를 하였다는 것이다. 데이터 정제 기법의 적용을 통해 상황 분류 기술의 발전에 도움이 될 것으로 예상된다. 또한, 제안된 기법은 다양한 음성 인식 및 분류 문제에 적용할 수 있어 다양한 분야에서의 응용 가능성이 높다.

향후 연구 방향은 다음과 같다. 더욱 효과적인 데이터 정제 방법을 도출하여 데이터 불균형 문제를 해결하고자 한다. 소수 클래스의 예측 성능과 모델의 일반화 성능을 개선함으로써 상황 분류 모델의 정확도를 더 높일 것이다.

음성 데이터 외에도 텍스트, 영상 등 다양한 유형의 데이터를 함께 활용하여 상황 분류 모델의 성능을 개선할 수 있다. 이를 통해 다양한 상황에서의 응용 가능성을 확장할 것이다.

참 고 문 헌

[1] Ye, Jiaxin, et al. "Temporal Modeling Matters: A Novel Temporal Emotional Modeling Approach for Speech Emotion Recognition." arXiv preprint arXiv:2211.08233 (2022).

[2] Osisanwo, F. Y., et al. "Supervised machine learning algorithms: classification and comparison." International Journal of Computer Trends and Technology (IJCTT) 48.3 (2017): 128-138.

[3] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." 2017 ieee international conference on acoustics, speech and signal processing (icassp). IEEE, 2017.

[4] Zhu, Xiaojin Jerry. "Semi-supervised learning literature survey." (2005).

[5] Likitha, M. S., et al. "Speech based human emotion recognition using MFCC." 2017 international conference on wireless communications, signal processing and networking (WiSPNET). IEEE, 2017.