wav2vec2.0을 활용한 한국어 음성 감정 분류를 위한 데이터 샘플링 전략

신미르[1], 신유현[2]
[1]인천대학교 컴퓨터공학부 학부생
[2]인천대학교 컴퓨터공학과 교수
tlsalfm820@inu.ac.kr, yhshin@inu.ac.kr

Data Sampling Strategy for Korean Speech Emotion Classification using wav2vec2.0

Mirr-Shin[1], Youhyun Shin[2]
[1]Dept. of Computer Science, Incheon National University
[2]Dept. of Computer Science, Incheon National University

요 약

음성 기반의 감정 분석은 인간의 감정을 정확하게 파악하는 데 중요한 연구 분야로 자리잡고 있다. 최근에는 wav2vec2.0과 같은 트랜스포머 기반의 모델이 음성 인식 분야에서 뛰어난 성능을 보이며 주목받고 있다. 본 연구에서는 wav2vec2.0 모델을 활용하여 한국어 감성 발화 데이터에 대한 감정 분류를 위한 데이터 샘플링 전략을 제안한다. 실험을 통해 한국어 음성 감성분석을 위해 학습 데이터를 활용할 때 감정별로 샘플링하여 데이터의 개수를 유사하게 하는 것이 성능 향상에 도움이 되며, 긴 음성 데이터부터 이용하는 것이 성능 향상에 도움이 됨을 보인다.

1. 서론

최근 딥러닝 기술의 발전에 따라 음성 인식 분야의 연구가 활발히 진행되고 있으며, 특히 음성 데이터를 활용한 감정 인식은 인간과 기계 간의 상호작용을 자연스럽게 만드는 핵심 요소로 간주된다. [1]은 wav2vec[2] 모델을 활용한 한국어 감정 분류 방법을 제안하였다. 본 연구에서는 wav2vec2.0[3]의 향상된 특징 추출 능력을 활용한 한국어 음성 감정 분류 모델을 제안한다. 본 연구에서는 제한된 레이블 데이터를 활용하여 효율적인 wav2vec2.0의 학습 방법을 위한 데이터 샘플링 방법에 대해 제안한다.

2. 관련 연구

2-1. wav2vec2.0

wav2vec은 Facebook AI에서 개발된 음성 인식을 위한 자기지도학습 모델이다. 이 모델은 크게 feature encoder와 context network로 이루어진다. Encoder는 입력값으로 받은 원시 오디오 데이터를 고차원의 특징으로 변환한다. Encoder는 원시 오디오의 짧은 시간 단위의 패턴을 학습하며, 이를 통해 음성의 기본적 특징을 추출한다.

wav2vec의 후속 모델인 wav2vec2.0에서는 여러 가지 구조적인 변화가 이루어졌다. 먼저, wav2vec의 causal CNN 기반 context network 부분이 트랜스포머 블록으로 교체되었다. 트랜스포머는 자연어 처리 분야에서 탁월한 성능을 보여주는 구조로, 이를 통해 wav2vec2.0은 오디오 데이터의 전체적인 문맥을 더욱 정교하게 파악하게 되었다. 또한, VQ-wav2vec 모델[4]에 적용된 양자화 모듈이 도입되어 원시 오디오 데이터의 표현을 더욱 압축하고 효율적으로 만들 수 있게 되었다. 이러한 아키텍처의 변화와 추가된 모듈 덕분에 wav2vec2.0은 원래의 wav2vec 모델보다 더욱 높은 성능과 효율성을 보인다.

3. 본론

wav2vec2.0 논문에서는 제한된 적은 수의 레이블 데이터만을 활용하여도 높은 성능을 달성할 수 있음을 보여주었다. 이러한 wav2vec2.0의 특성을 바탕으로 본 연구에서는 AI-Hub에서 공개된 한국어 음성 데이터를 활용하여 실험을 진행하였다. 데이터 제한 방법으로는 두 가지를 사용했다. 첫 번째는 데이터 양을 기반으로 한 제한, 두 번째는 음성 데이터의 길이를 기반으로 한 제한이다. 이러한 두 가지 방법을 통해 실험을 진행하며, 각 방법에 따른 성능의 변화를 관찰하였다. 실험에서의 성능 변화를 통

해 최적의 데이터 제한 방법을 찾는 것이 목적이다.

4. 실험

실험에 사용한 GPU는 NVIDIA A100 PCIe 40GB GPU를 사용하였다. batch size는 32, epoch은 10으로 통일하여 실험을 진행하였다. 사용한 모델은 huggingface의 facebook/wav2vec2-large-robust 모델을 사용하여 실험을 진행하였다.

4-1. 데이터 셋

본 연구에서는 AI-Hub에 공개된 '감성 및 발화 스타일별 음성합성 데이터'를 사용하여 실험을 진행하였다.[1] 7가지 감정(기쁨, 슬픔, 분노, 불안, 상처, 당황, 중립)에 대해 분류하는 실험을 진행하였다.

4-2. 데이터 샘플링 전략

4-2-1. 데이터 양에 따른 성능 변화

<표 1>은 감정별 데이터의 샘플링 양을 조절하여 성능 변화를 관찰한 결과이다. 총 7개의 감정에 대해서 데이터를 샘플링하여 성능 변화를 살펴보았다.

<표 1> 데이터 샘플링 양 증가에 따른 모델 정확도

데이터 수(개)	5000	10000	15000	20000	25000	30000
Acc.(%)	79.20	86.32	89.33	92.34	92.06	92.79

<표 1>에서 알 수 있듯이, 데이터의 수가 많아질수록 성능이 향상되는 경향이 있음을 알 수 있다. 특히나 감정별로 5,000개씩만 샘플링할 경우보다 10,000개씩 샘플링할 경우 79.20에서 86.32로 7.12의 정확도가 향상되었다. 25,000개의 경우 길이가 고려되지 않고 랜덤하게 데이터가 샘플링되어 성능이 하락한 것으로 보여진다.

4-2-2. 길이 기반 샘플링 전략

<표 2>는 음성 데이터의 길이를 기준으로 긴 데이터부터 순차적으로 샘플링할 경우와 짧은 데이터부터 샘플링할 경우에 대한 실험을 진행하였다.

<표 2> 데이터 길이에 따른 모델 정확도

데이터 수(개)	5000	10000	15000	20000	25000	30000
Random	79.20	86.32	89.33	92.34	92.06	92.79
Short	78.49	85.89	88.93	89.82	91.31	92.25
Long	82.90	87.20	89.51	92.41	92.46	93.70

<표 2>에서 알 수 있듯이, 긴 음성 데이터부터 순차적으로 학습시킬 경우(Long), 짧은 음성 데이터부터 학습시켰을 때(Short)보다 성능이 향상되었다. 모든 경우에서 랜덤한 샘플링(Random)의 경우보다 데이터 길이가 긴 것부터 샘플링하여 학습할 경우 정확도 향상을 보였다. 또한 데이터 샘플링 양이 많아질수록 성능이 향상됨을 알 수 있었다.

5. 결론

본 논문에서는 wav2vec2.0을 활용하여 한국어 음성 감정 분류 모델을 학습하기 위한 데이터 샘플링 방법에 대해 제안하였다. 실험 결과를 통해 학습 데이터의 양이 증가할수록 정확도가 향상되는 경향이 있음을 보였다. 그러나 데이터 양이 증가할수록 학습에 소요되는 시간 또한 증가하기 때문에 제한된 양질의 데이터를 사용하여 학습하는 것이 중요하다. 따라서 본 논문에서는 랜덤하게 데이터를 샘플링할 때보다, 길이가 긴 음성 데이터부터 샘플링하여 데이터의 양을 증가시킬 경우 가장 큰 성능 향상이 있음을 보였다.

사사문구

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 학석사연계ICT핵심인재양성사업의 연구결과로 수행되었음" (IITP-2023-RS-2023-00260175)

참고문헌

[1] 안영도, 한상욱, 이성주, & 신종원. Wav2vec 특징 기반의 한국어 음성감정인식. 한국통신학회 학술대회논문집, 2021, 11-12.

[2] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862.

[3] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[4] Baevski, A., Schneider, S., & Auli, M. (2019). vq-wav2vec: Self-supervised learning of discrete speech representations. arXiv preprint arXiv:1910.05453.

1) https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=466