하이패스필터를 적용한 ResNet 기반의 새소리 탐지

서강대학교 | 컴퓨터공학과 | Bigdata Processing & DB LAB
이강우, 권태현, 가상민, 김윤영, 정성원

- 2022. 8.26. -

목 차

개 요

이미지 출처 : kaggle BirdCLEF 2022

- **새소리: 생태계 환경 지표**

효율적인 새소리 식별 기법 개발

(by Kaggle)

개 요

- **데이터 셋의 문제점**

1. Weakly-Labeled

'파일단위'로 어떤 새가 울었는 지에 대한 정보가 주어져, 정확히 어느 구간에서 울었는지 알 수 없음

→ **파일 내 'Call', 'Nocall' 분류 필요**

2. 도메인 불일치

녹음환경(기기, 잡음 등)의 차이로, validation 및 test set에서의 성능저하

→ **새의 고유 주파수 추출 필요**

3. 데이터 분포의 불균형

녹음이 힘든 환경에 서식하거나, 새가 희귀종인 경우 등 새 종별 제공된 데이터 양에 차이가 있음

→ **데이터 증강기법 활용 필요**

☞ **오디오를 분석하기 위해 이미지로의 변환 필요**

☞ **이미지 학습 및 분류를 위해 CNN 모델 필요**

관련연구

- **2021년도 동일 Task 상위팀 분석**
  - (공통) Pre-trained CNN 사용하여 새 소리 분류

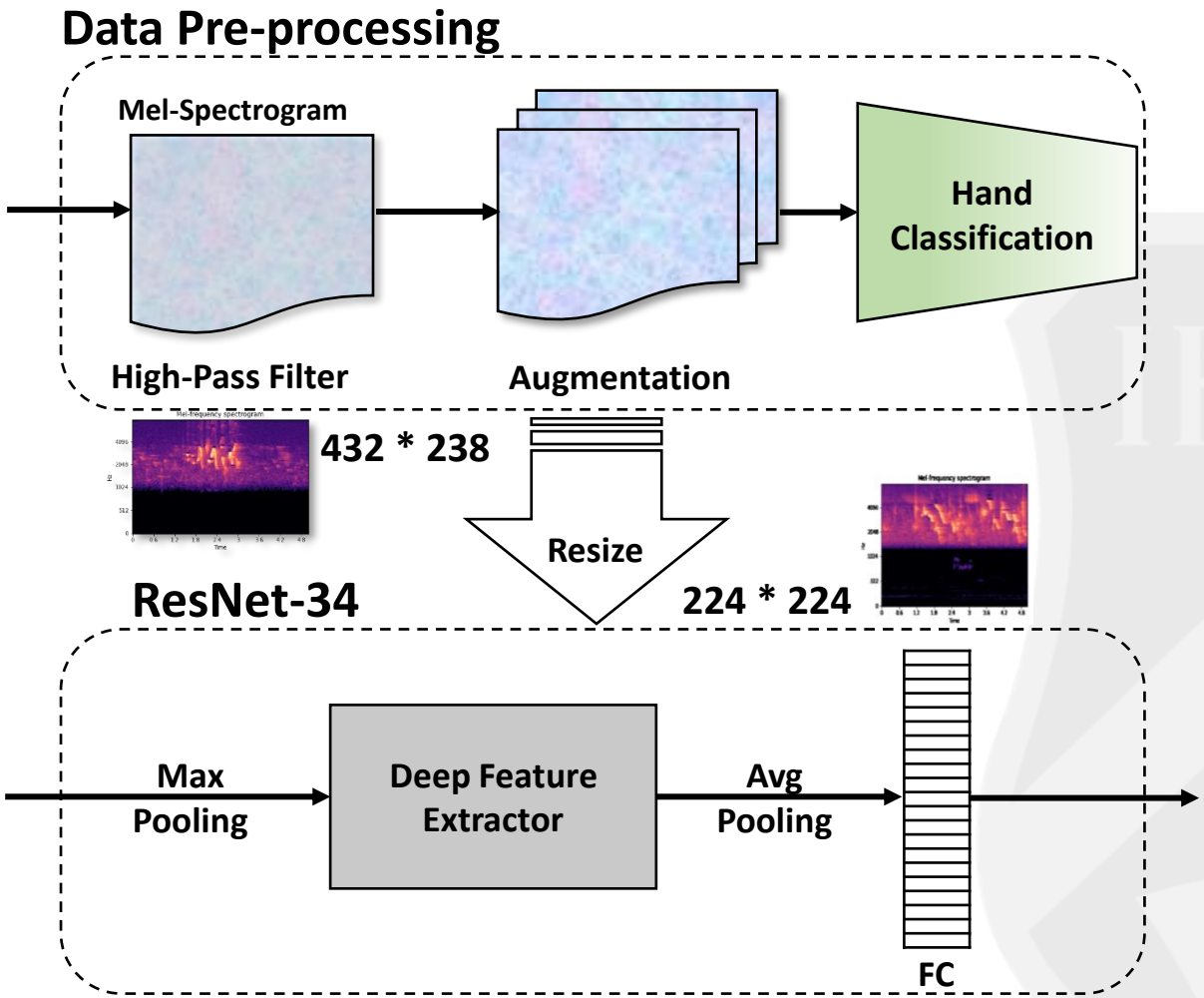
순위	아이디어	효과
1위[1]	Nocall Detector	Weakly-Labeled 보완
10위[2]	데이터 증강	데이터 분포의 불균형 보완
12위[3]	훈련 데이터에 잡음을 섞음	도메인 불일치 보완

⇒ **새 울음 소리가 "음성"이라는 점을 간과**

⇒ **단순 이미지 변환 후 모델 학습**

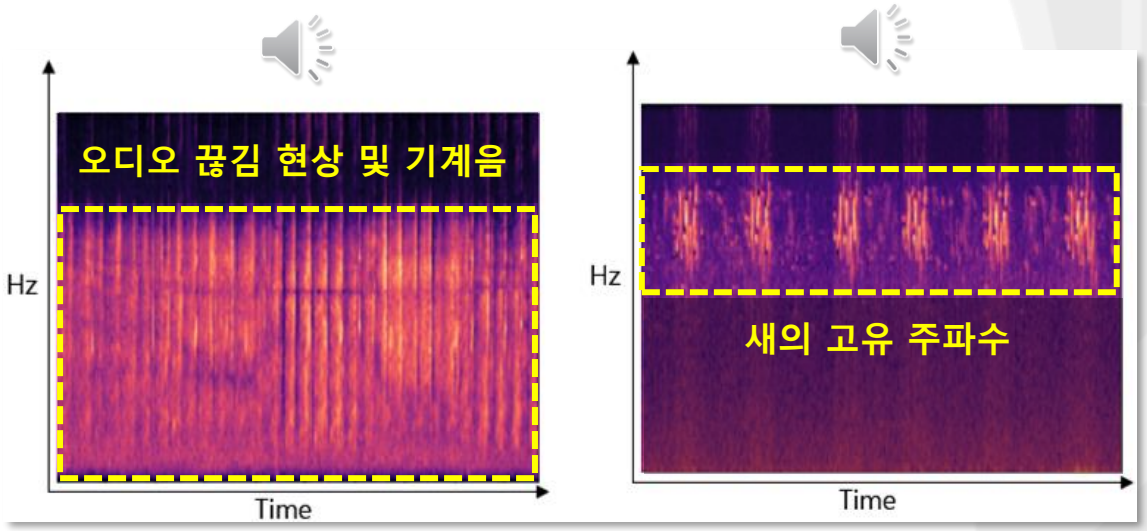
* Reference
[1] Naoki Murakami, Hajime Tanaka and Masataka Nishimori, "Birdcall Identification Using CNN and Gradient Boosting Decision Trees with Weak and Noisy Supervision", in: CLEF Working Notes 2021
[2] M. V. Conde, N. D. Movva, P. Agnihotri, S. Bessenyei, K. Shubham, "Weakly-Supervised Classification and Detection of Bird Sounds in the Wild. A BirdCLEF 2021 Solution", in: CLEF Working Notes 2021
[3] Jan Schlüter, "Learning to Monitor Birdcalls From Weakly-Labeled Focused Recordings", in: CLEF Working Notes 2021

모델 Block Diagram



데이터 전처리 과정

- **Data Selection**
  - 새의 고유 주파수 특성을 확인하기 위한 'Rating' 필터링
  - 'Rating 3.5' 미만인 데이터는 오디오 끊김 현상 및 기계음 다수 포함
  - 'Rating 3.5' 이상인 데이터는 새의 고유 주파수 특성 확인 가능

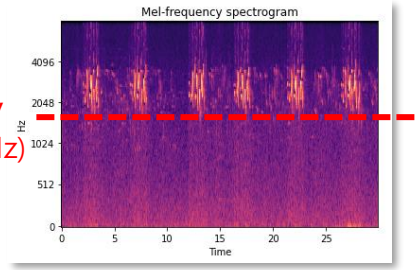


'Rating 0'인 오디오 'Rating 5'인 오디오

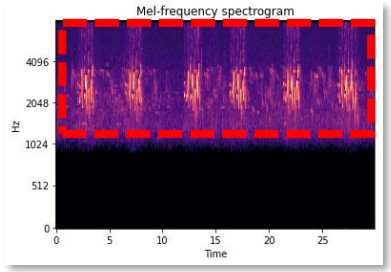
데이터 전처리 과정

- **Min-Frequency 추출과 하이패스필터 적용**
  - Mel-Spectrogram을 직접 분석하여 새 종별 Min-Frequency 추출
  - Min-Frequency를 통해 하이패스필터를 적용 새 종별 고유 주파수 확인 및 '도메인 불일치' 해결

Min-Frequency
(e.g. Akiapo : 1,300Hz)



하이패스필터 적용

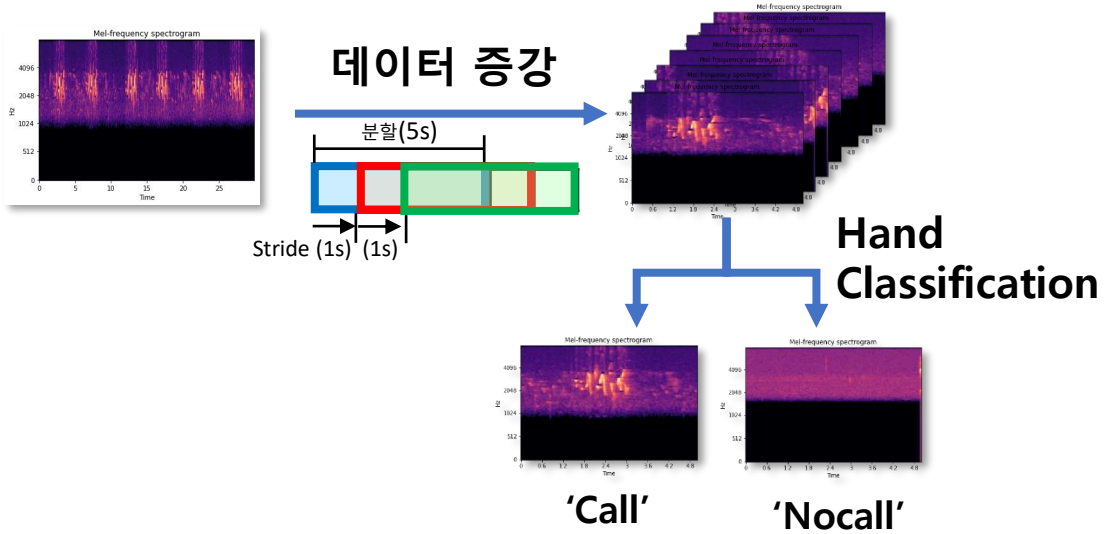


새 종별 Min-Frequency 추출 결과

새 종	Min-Freq	새 종	Min-Freq	새 종	Min-Freq
Akiapo	1,300	Hawama	1,950	Jabwar	980
Aniani	2,100	Hawcre	1,550	Maupar	1,500
Apapan	1,700	Hawgoo	350	Omao	940
Barpet	0	Hawhaw	1,100	Puaioh	1,800
Crehon	0	Hawpet1	2,100	Skylar	1,600
Elepai	1,500	Houfin	1,950	Warwhel	2,240
Ercfra	400	Iiwi	1,800	Yefcan	1,750

데이터 전처리 과정

- **데이터 증강과 Hand Classification**
  - Kaggle 테스트 데이터와 동일하게 5초 단위로 분할
  - 데이터 증강을 위해 1초 단위의 Stride 적용
  - 기존 1,265개 → 36,294개(약 28배 증가)로 데이터 증강
  - 증강된 데이터를 통해 1,197개의 'Nocall' 분류로 'Weakly Labeled' 해결 및 '데이터 분포의 불균형' 일부 해결



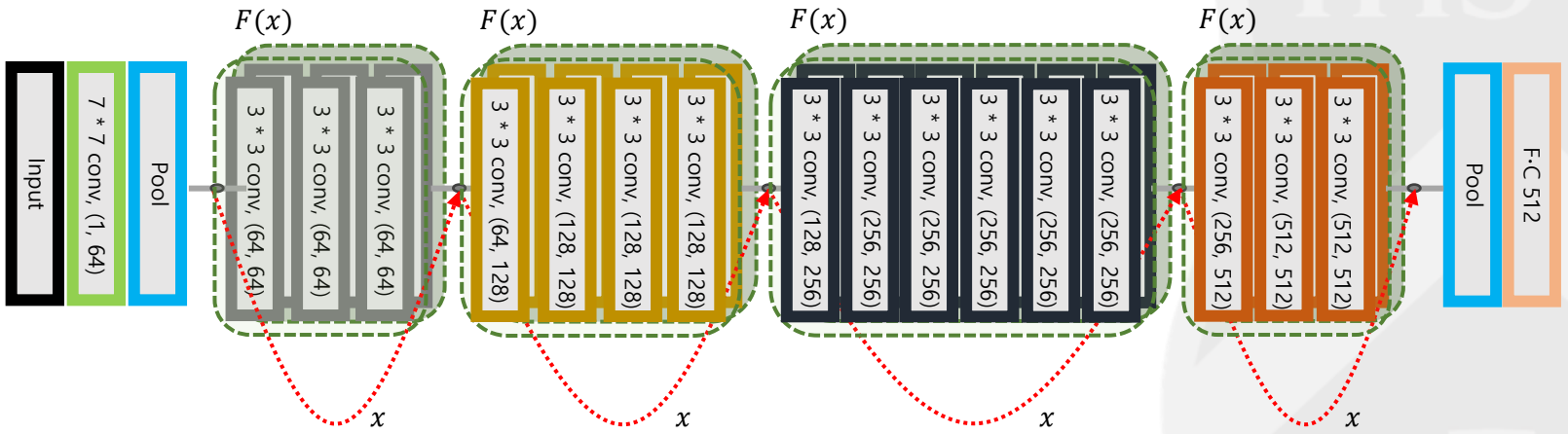
데이터 증강 결과 예시

새 종	기 존	증강 후
Apapan	47	**1,722**
Hawcre	20	**1,036**
Houfin	322	**2,776**
Iiwi	37	**1,550**
Jabwar	78	**3,331**
Warwhel	71	**2,030**
총 계 : 기존 1,265개 증강 후 36,294개(약 28배 증가)		

분류 모델

- **ResNet-34[4]**
  - Mel-Spectrogram 분류는 적절한 파라미터와 Layer의 개수로도 충분한 성능 확인
  - 입력데이터(224*224) → Convolution Layer(7*7 Kernal) → Hidden Layer(32개) → Fully-Connected Layer(21종 분류)

ResNet-34 Diagram



* Reference
[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. "Deep Residual Learning for Image Recognition", Proc. CVPR 2016

실험 결과

- (A) + (B) + (C)이 Best F1-Score
- (B) 및 (C)의 F1-Score가 낮은 이유는 새소리 이외의 부분이 과적합된 것으로 분석

⇒ **하이패스필터**가 가장 **유의미한 성능향상** 결과를 보여줌

전처리 방식에 따른 F1-Score 결과

Index	Case	F1-Score
1	하이패스필터 (A)	0.95
2	데이터 증강 (B)	0.61
3	Nocall Hand Classification (C)	0.59
4	(A) + (B)	0.96
5	(A) + (C)	0.96
6	(B) + (C)	0.64
7	**(A) + (B) + (C)**	**0.97**
8	전처리 미적용	0.73

결 론

☑ 결론 및 향후 연구

- **이미지 모델(ResNet-34)을 통한 음성분류에 대한 전처리 기법 제안**
  - Min-Frequency 추출 및 하이패스필터 적용
  - Hand Classification
  - 데이터 증강

    ⇒ **하이패스필터**가 가장 **유의미한 성능향상** 결과를 보여줌

- 다른 이미지 분류모델에서 위의 제안된 전처리 기법이 동일한 성능향상을 나타내는지 추가연구 필요

감사합니다.