하이패스필터를 적용한 ResNet 기반의 새소리 탐지

이강우 권태현 가상민 김윤영 정성원†
서강대학교 컴퓨터공학과
lkw559@naver.com, sungro96@sogang.ac.kr, sm3minka@sogang.ac.kr, aff1105@sogang.ac.kr, jungsung@sogang.ac.kr

ResNet-based Birdsong Detection with High-pass Filter

Kangwoo Lee, Taehyeon Kwon, Sangmin Ga, YunYeong Kim, Sungwon Jung
Department of Computer Science and Engineering, Sogang University

요 약

BirdCLEF 2022(Kaggle)은 새의 울음소리를 듣고, 21종의 새 종을 판별하는 대회이다. 총 5,500개로 구성된 1분 길이의 테스트 데이터가 주어졌을 때, 21종의 새에 대하여 5초 단위로 'Call(True)' 또는 'Nocall(False)'를 분류하는 것을 목표로 한다. 본 논문에서는 Weakly Labeled 등 BirdCLEF 2022에서 제공하는 데이터의 세 가지 문제점을 제시하고, 해결 방안으로는 데이터 증강, Min-Frequency 추출, 하이패스필터 적용 및 'Nocall' 제거를 위한 Hand Classification 등 총 4단계의 데이터 전처리를 진행하였다. 이후 이미지 분류에 좋은 성능을 보여주는 ResNet-34 모델을 구축하여 전처리 된 데이터셋을 학습하였고, F1 score를 계산한 결과 97%의 성능을 달성했다.

1. 서 론

새의 생존 여부는 생태계의 오염도와 변화량을 확인할 수 있는 지표 중 하나로, 지속적인 관찰이 필요하다. 그러나 이들의 상태를 직접 관찰하고 판단하는 것은 많은 시간이 요구된다. 특히 희귀종의 경우, 개체 수가 적어 정확한 판단을 내리기에 어려운 실정이다. 이에 따라 Kaggle[1]은 매년 BirdCLEF[2]를 개최하여, 데이터가 적은 환경에서 새소리를 식별하는 기법을 개발하고, 생태계의 변화를 파악하는데 기여한다.

BirdCLEF 2022에서 제공하는 데이터는 세 가지 문제점이 있다. 첫 번째는 Weakly-Labeled 문제다. 제공된 데이터는 파일 단위로 Labeling 되어 오디오 파일의 어느 구간에 새가 울었는지 정확히 파악할 수 없다. 두 번째는 파일별 녹음환경(기기, 잡음 등)의 차이로 인한 도메인 불일치 문제가 있다. 마지막으로 데이터셋의 새 종별 데이터 분포의 불균형으로 데이터가 적은 새 종의 학습이 원활하지 않은 문제가 있다. 따라서 본 논문에서는 이러한 세 가지 문제를 해결하고자 네 단계의 전처리를 수행한다. Weakly-Labeled 문제를 해결하고 하이패스 필터를 적용하기 위한 새 종별 Min-Frequency 추출을 위해 음성데이터를 멜-스펙트로그램으로 변환한다. 이후 파라미터 및 Layer 개수 대비 좋은 성능을 보여주는 ResNet[3]을 사용하여 학습을 진행한다.

본 논문의 구성은 서론에 이어 다음과 같다. 2장은 지난 BirdCLEF에 참여한 상위 팀의 접근 방법을 논의한다. 3장은 대회에서 제공된 데이터셋의 전처리 과정을 설명하고, ResNet 모델 구조와 학습 과정을 설명한다. 4장에서는 이에 따른 실험 결과를 설명하고, 마지막 5장에서 추후 연구 목표와 함께 논문의 결론을 맺는다.

2. 관련연구

BirdCLEF 2021에서 제공된 데이터는 Weakly-Labeled, 도메인 불일치 문제 등이 있다. 이러한 문제를 해결하고자 지난 BirdCLEF 2021에 참여한 팀을 분석했다. 상위권에 있는 많은 팀은 Pre-trained CNN을 활용하였으며, 그중에서도 우승팀[4]은 Weakly-Labeled 문제를 해결하고자 Nocall Detection을 수행하였다. 또한 12위팀[5]은 도메인 불일치문제를 해결하기 위해 데이터셋에 잡음을 섞거나, 고주파를 낮추는 등 데이터를 전처리했다. 마지막으로 10위팀[6]은 데이터 불균형 문제를 해결하기 위해 데이터 증강을 수행했다.

대부분의 상위 팀들은 새소리 데이터가 음성이라는 점을 간과하고 단순 이미지로 변환 후 모델에 학습시켰다는 아쉬움이 있었다. 따라서 본 논문에서는 지난 대회에서 상위 팀들의 장점을 취하면서도, 새소리가 음성이라는 점에 집중하여 세 가지 문제점을 해결하였다. 첫째, Hand Classification을 통해 Weakly-Labeled 문제를 해결하였다. 둘째, 도메인 불일치문제를 해결하기 위해 Min-Frequency를 적용하였으며, 마지막으로 데이터 분포의 불균형 문제를 해결하고자 데이터 증강을 수행했다. 이에 관한 내용은 3장에서 설명한다.

3. 하이패스필터를 적용한 ResNet 기반 새소리 탐지

3.1. 데이터 전처리 과정

[그림 1]은 제안한 모델의 데이터 전처리 개략도를 나타낸 것이다. ① 입력데이터는 잡음이 적은 새소리를 학습시키기 위해 오디오 녹음 품질 정보를 제공하는 'Rating'이 '3.5' 이상인 오디오 파일을 추출한다. ② 이 데이터를 멜-스펙트로

그램으로 변환하여 Hand labeling으로 새 종별 Min-Frequency를 확인하고, ③ 멜-스펙트로그램으로 변환된 Raw 오디오에 하이패스필터를 적용한다. ④ 하이패스필터가 적용된 오디오를 테스트 데이터와 동일하게 5초 단위로 나누며, 데이터 증강을 위해 1초 단위의 Stride를 적용한다. ⑤ 증강된 데이터들을 Hand Classification을 통해 'Call'과 'Nocall'로 분류한다.

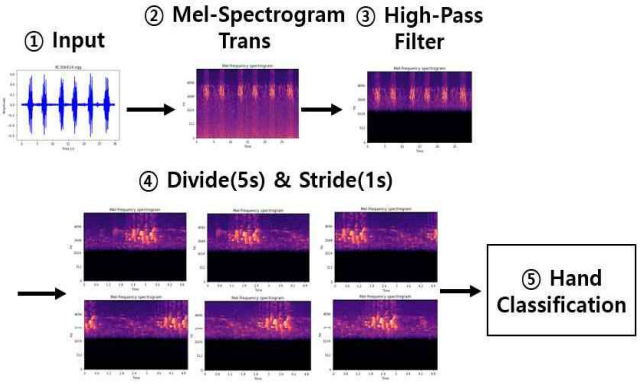
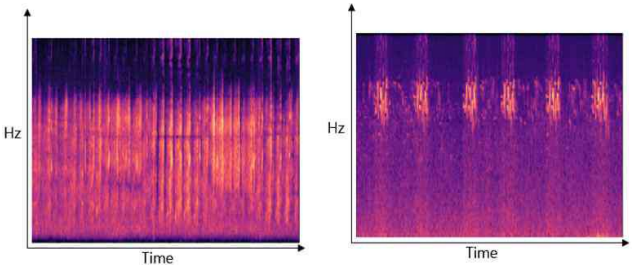


그림 1. 데이터 전처리 개략도

3.1.1. Data Selection

훈련 데이터를 선별하기 위해, 'Rating'이 '3.5' 이상인 오디오를 필터링한다. 이유는 [그림 2. (a)]에서와 같이 'Rating'이 '3.5' 미만의 오디오는 끊김 현상이나 기계음 또는 잡음(바람 소리) 등이 포함되어 새소리에 대한 고유 주파수 특성을 확인하기 어렵기 때문이다. 반대로 [그림 2. (b)]은 'Rating'이 '5'인 오디의 멜-스펙트로그램이다. 'Rating'이 낮은 오디오보다 적은 잡음이 포함되어 새소리의 고유한 주파수 특성이 분명하므로 'Rating'이 '3.5' 이상인 오디오를 추출하여 학습시킨다.



(a) 'Rating'이 '0' (b)'Rating'이 '5'

그림 2. 'Rating'별 멜-스펙트로그램 비교

3.1.2. Min-Frequency 추출

[표 1]은 오디오에 포함된 소음 등을 최소화하고 새소리의 고유 주파수 특성을 추출하기 위해 직접 청취하고 분석하여 Min-Frequency를 추출한 결과이다. 이를 통해 오디오에 새 종별 고유 주파수의 Min-Frequency를 확인하고, 오디오에 하이패스필터를 적용한다.

표 1. 새 종별 Min-Frequency 확인 결과

새 종	Min-Fre	새 종	Min-Fre	새 종	Min-Fre
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	Warwhe1	2,240
Ercfra	400	Iiwi	1,800	Yefcan	1,750

3.1.3. 데이터 증강과 Hand Classification

테스트 데이터와 동일하게 오디오를 5초 단위로 나누고 1초 단위의 Stride를 적용하여 데이터 증강을 수행한다. [표 2]는 새소리별 증강된 데이터 개수의 예시이다. 데이터 증강 결과 기존 1,265개에서 36,294개로 2.8배 증가하였다. 이후 Hand Classification을 통해 1,197개의 'Nocall'을 분류하여 Weakly-Labeled 문제를 해결하였다.

표 2. 데이터 분할 및 증강 결과 예시

새 종	기존	증강 후	새 종	기존	증강 후
Apapan	47	1,722	Iiwi	37	1,550
Hawcre	20	1,036	Jabwar	78	3,331
Houfin	322	2,776	Warwhe1	71	2,030
총 계 : 기존 1,265개 / 증강 후 36,294개(2,869% 증가)					

3.2. 분류모델

본 논문에서는 분류모델을 ResNet-34를 적용하였다. 그 이유는 첫째, Weakly-Labeled 문제를 해결하고 하이패스필터를 적용하기 위한 새 종별 Min-Frequency를 추출하기 위해 음성데이터를 이미지(멜-스펙트로그램)로의 변환이 필수적이었고, 이미지 분류에서 좋은 성능을 보여주는 CNN 모델의 적용이 필요하기 때문이다. 둘째, 멜-스펙트로그램을 활용한 이미지 분류 문제는 'ImageNet 프로젝트'와 같은 복잡한 대회가 아니므로 적절한 파라미터와 Layer의 개수로도 충분한 성능을 보여주는 모델은 ResNet-34로 판단하였기 때문이다.

학습할 입력데이터는 432*238의 픽셀을 가지고 있으며, ResNet-34을 사용하기 위해 이미지의 크기를 224*224의 이미지로 조정하였다.

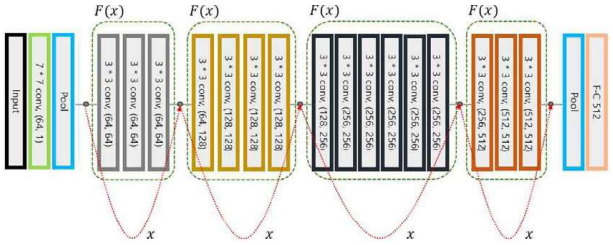


그림 3. 새소리 식별에 쓰인 ResNet34 Block Diagram

모델의 구성은 [그림 3]과 같다. 입력데이터는 7*7 커널을 적용한 Convolution Layer와 32개의 중간 Layer를 통과한다. 중간 Layer 사이에 Skip-Connection($f(x)+x$)을 적용하며, 마지막 Layer에서는 Fully-Connected Layer를 통해 분류학습을 진행한다.[3] 또한 모델학습 과정에서 파라미터 조정을 위하여 Batch Size와 Learning Rate의 비교실험을 진행하였으며, 그 결과 Batch Size가 64, Learning Rate는 0.0001일 때, 가장 낮은 Loss를 확인할 수 있었다.

4. 실 험

데이터는 총 1,121개의 오디오가 21개의 클래스로 구성되어 있으며, 클래스별 8:2 비율로 나누어 학습데이터와 검증 데이터로 사용하였다. 본 논문에서 제안된 하이패스필터, 데이터 증강, 'Nocall Hand Classification'의 성능을 비교하기 위해 Batch Size와 같은 하이퍼 파라미터는 모두 고정하고, 20 Epoch을 반복하였으며, 성능평가 지표로는 F1 score를 사용했다.

실험 결과는 [표 3]과 같이 전처리를 적용하지 않은 데이터로 학습시킨 모델에 비해 하이패스필터, 데이터 증강, 'Nocall Hand Classification'을 모두 적용한 데이터로 학습시킨 모델 (A)+(B)+(C)이 '0.97'로 더 좋은 성능을 보여주었으며, 전반적으로 하이패스필터 적용이 가장 유의미한 성능향상을 보여주었다.

표 3. 학습 데이터에 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

또한, 전처리하지 않은 데이터로 학습시킨 모델보다 데이터 증강 (B), 'Nocall Hand Classification (C)'을 적용한 데이터로 학습시킨 모델의 F1-score가 낮은 것을 확인할 수 있었는데, 이는 모델이 학습시켜야 하는 새소리 이외의 부분이 과적합 된 것으로 분석하였다.

5. 결 론

본 논문에서는 Weakly Labeled, 도메인 불일치, 데이터 분포 불균형의 문제를 가지고 있는 '음성' 데이터에 대한 전처리 방법을 제시하였다. 그중 하이패스필터를 적용한 모델에서 30%의 성능향상이 있었지만, 데이터 증강과 'Nocall Hand Classification'을 적용한 모델은 전처리하지 않은 모델보다 성능이 떨어지는 것을 확인할 수 있었다. 추후 이에 대한 정확한 분석이 필요하며 ResNet 이외의 다른 모델에서 논문에서 제시된 전처리 적용 시, 동일한 성능향상의 결과가 나오는지에 대해 추가연구가 필요하다.

하이패스필터 적용을 위한 개체별 Min-Frequency 추출은 시간이 소요되는 작업이다. 하지만 하이패스필터의 성능 향상률을 보았을 때, 추후 개체 식별을 위한 오디오 분류 시 하이패스필터를 통한 처리방안 연구에 활용할 수 있을 것으로 기대한다.

참 고 문 헌

[1] Kaggle, https://www.kaggle.com/

[2] BirdCLEF 2022, https://www.kaggle.com/competitions/birdclef-2022

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun. "Deep Residual Learning for Image Recognition", Proc. CVPR 2016, pp.770-778, 2016

[4] 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, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.

[5] Jan Schlüter, "Learning to Monitor Birdcalls From Weakly-Labeled Focused Recordings", in: CLEF Working Notes 2021, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.

[6] 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, CLEF: Conference and Labs of the Evaluation Forum, Sep. 2021, Bucharest, Romania, 2021.