PNU SW학습공동체 중간보고서

team 33

1. 프로젝트 소개

배경 및 필요성

조류는 생태계의 건강을 측정하는 데 있어 매우 중요한 지표 생물이다. 조류의 울음소리는 종 식별, 개체 수 파악, 이동 경로 추적 등 다양한 생태학적 연구에 활용되고 있다. 그러나 전문적인 조류 식별 능력을 가진 인력이 부족하고, 수작업 기반의 관측 방식은 시간과 비용이 비효율적으로 많이 들어가며 주관적인 오류 가능성도 크다.

이에 따라 최근에는 조류 소리를 인공지능으로 자동 분류하는 기술이 주목받고 있다. 특히 BirdCLEF와 같은 글로벌 대회는 조류 음성 인식 기술의 발전을 이끄는 촉매 역할을 하고 있으며, 실시간으로 생태계 모니터링 및 보호를 위한 기초적인 기술로써의 필요성이 커짐에 따라 이에 대한 개발의 필요성이 커지고 있다.

개발목표 및 주요내용, 세부내용

이번 task에서의 최종 목표는 AI 모델을 활용하여 조류의 울음소리를 정확히 분류할 수 있는 시스템을 구축하는 것이다. 이를 위해 다음과 같은 세부 내용을 중심으로 개발을 진행했다.

주요 내용

1. 다양한 환경에서 녹음된 조류 음성 데이터(wav, ogg)를 분류 가능한 형태로 처리한다.

2. 사전학습된 모델을 기반으로 성능 높은 분류 모델을 구현한다.

3. 모델 활용을 위한 오디오에서 이미지로 변환하는 파이프라인을 구축한다.

4. Macro F1-score, Accuracy 등의 평가 지표를 기반으로 성능을 자체 검증한다.

5. 최종적으로 완성된 모델을 kaggle에 제출하여 성능을 평가받는다.

사회적가치 도입 계획

본 프로젝트는 기술 개발을 넘어 다음과 같은 사회적 가치 실현을 목표로 두고 있다.

1. 환경 보호 및 생물 다양성 보존 기여

조류 분류 자동화를 통해 전문가 부족 문제를 보완하고, 광범위한 생태계 데이터를 빠르고 정확하게 분석함으로써 멸종 위기 조류와 같은 종들의 실시간 탐지 및 보호에 기여할 수 있다.

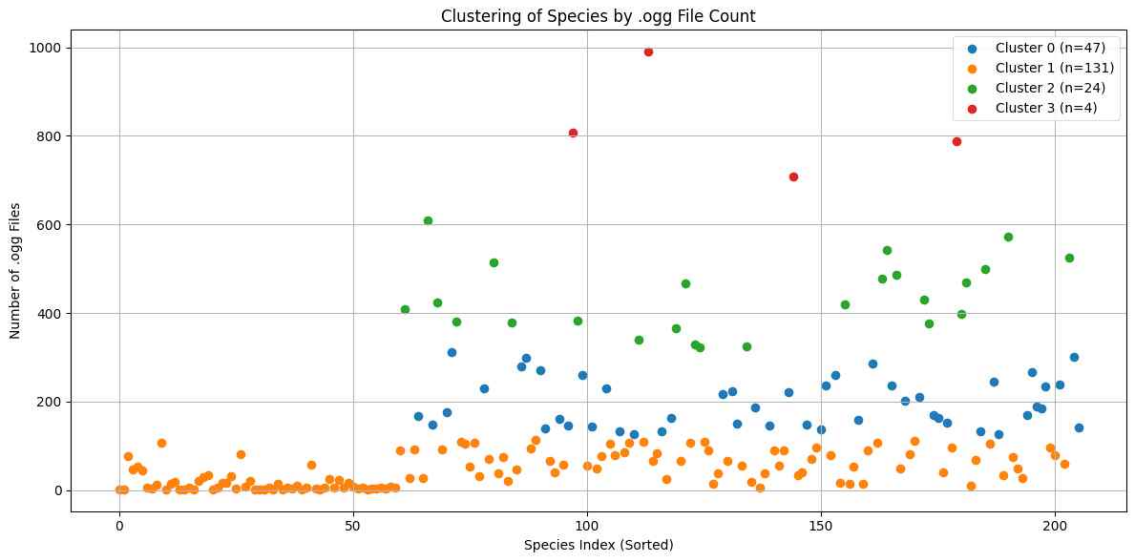
2. Citizen Science의 확장 기반 마련

조류 울음소리를 스마트폰으로 녹음해 자동으로 종을 식별해 주는 앱으로 확장이 가능하며, 전문가뿐 아니라 일반 대중의 참여를 유도하여 자연 보호 인식 제고에 기여 할 수 있다.

3. AI 기술의 친환경적 활용 사례 제시

이번 프로젝트가 성공적으로 마무리된다면 인공지능 기술을 환경 보존에 접목한 대표 사례로 인식될 수 있으며, AI의 지속 가능한 활용 모델로서의 의미가 있을 것이다.

2. 상세설계



모델 학습에 앞서, 전체 클래스별 학습 데이터 수의 분포를 파악하기 위해 클러스터링을 기반으로 시각화를 수행하였다.

각 클래스에 포함된 .ogg 파일 수를 기준으로 K-Means 클러스터링을 적용한 결과,
클래스 간의 샘플 수 차이가 최대 약 1,000개 이상으로, 매우 심각한 수준의 데이터 불균형이 존재함을 확인할 수 있었다.
아래 그래프는 클래스별 샘플 수를 시각화한 것으로,
주황색(Cluster 1)은 데이터가 적은 클래스, 초록색(Cluster 2)과 빨간색(Cluster 3)은 상대적으로 많은 클래스를 나타낸다.

이러한 불균형 문제를 완화하기 위해 다음과 같은 방법들이 적용될 수 있다.

가중치 손실 함수 적용 (Loss Weighting)
클래스별 데이터 수에 반비례하는 가중치를 손실 함수에 부여하여,
데이터가 적은 클래스의 오차를 더 크게 반영하도록 학습을 유도한다.

데이터 증강 (Data Augmentation)
음성 데이터에 다양한 변형을 가해 학습 샘플을 인위적으로 증가시키는 방법으로,
Pitch shifting, noise injection 등의 기법을 활용할 수 있다.

Focal Loss 활용
자주 맞추는 쉬운 샘플에는 손실을 줄이고, 희귀하고 어려운 샘플에는 손실을 증폭시켜 학습을 집중시키는 방식이다. 분류 불균형에 강건하며, 성능 향상에 기여할 수 있다.

새소리 오디오 신호 전처리

1. Bandpass filter

기술 설명

밴드패스 필터(Bandpass filter) 는 하한주파수와 상한 주파수를 설정하여, 그보다 낮거나 높은 주파수를 제거하거나 감쇠하여 특정 주파수 대역만을 통과시키는 필터이다. 오디오 신호 처리에서는 원하는 신호 주파수 영역을 선택적으로 추출하고, 불필요한 노이즈를 효과적으로 제거하는 데 사용된다.

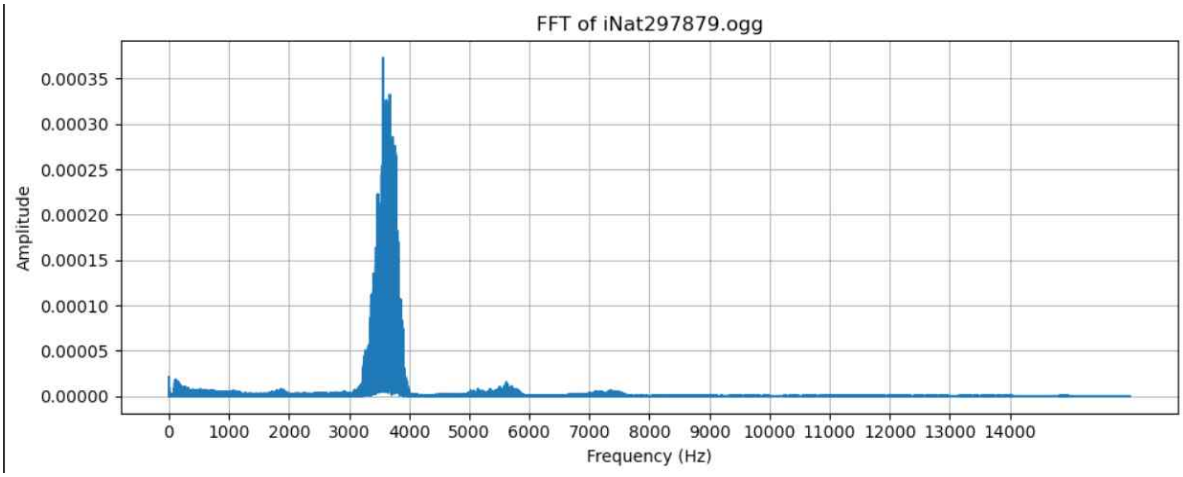
사용 목적

조류의 음성 신호는 일반적으로 특정 주파수 범위에 집중되어 있다.

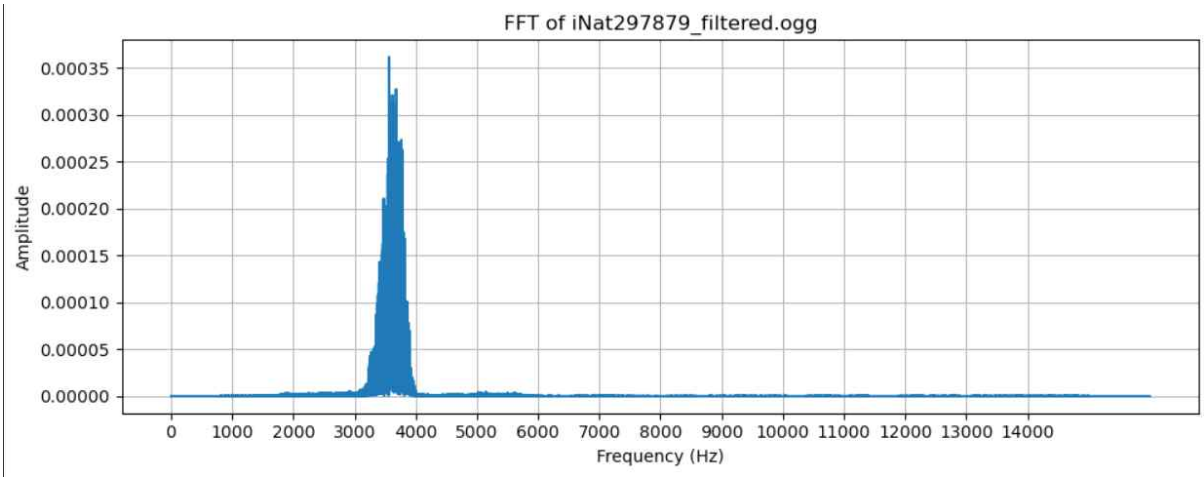
밴드패스필터를 적용하여 이 특정 주파수 대역을 종별로 선택적으로 추출함으로써, 분석하고자 하는 핵심 신호의 명확성을 극대화하고자 했다.

또한 필터링으로 전처리된 깨끗한 신호는 이후 MFCC와 같은 음향적 특징 추출 과정의 품질을 높여주고 이는 최종적으로 종 분류 및 식별을 위한 모델의 성능 향상과 연결되기에 이 기술을 사용하였다.

적용 예시



<필터링 전>



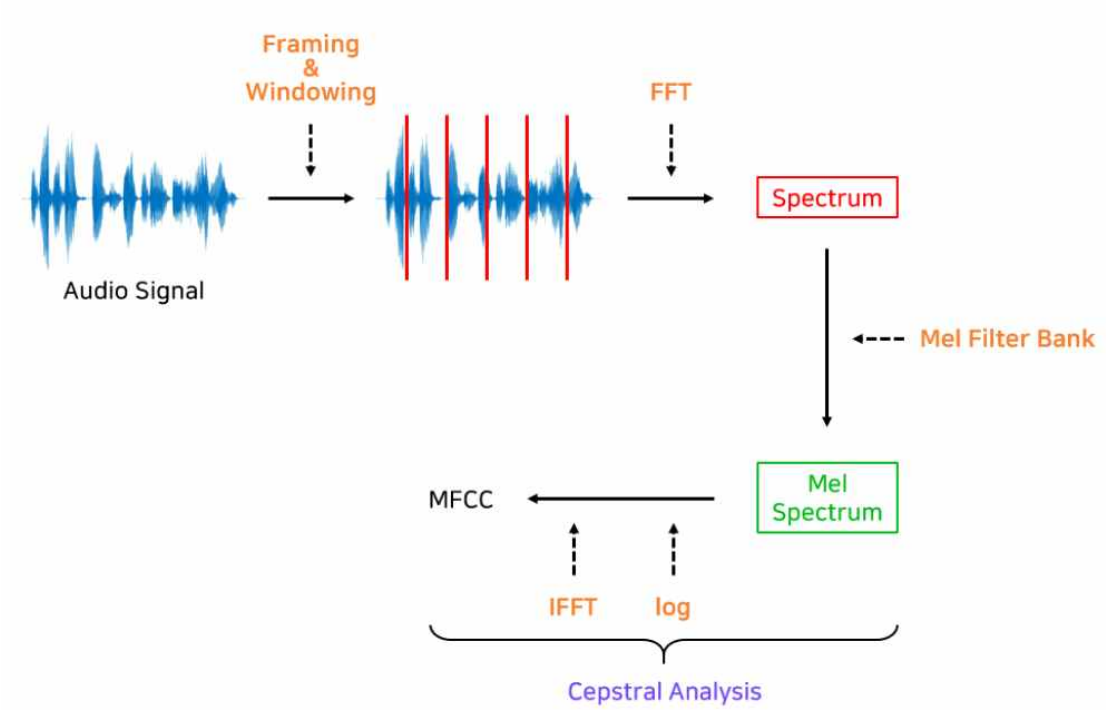
<필터링 후>

2. Mfcc 추출

기술 설명

MFCC(Mel Frequency Cepstral Coefficients)는 오디오 신호 처리 분야에서 주로 사용되는 음향적 특징의 일종으로, 음성 및 오디오 신호를 효과적으로 표현하기 위한 기술이다.

MFCC는 사람의 청각 인지 시스템의 특성을 모방하여 개발된 특징으로, 음성 및 조류 분류 등 다양한 오디오 인식 분야에서 널리 사용된다.



<추출 원리>(https://brightwon.tistory.com/11)

사용 목적

Train_audio 데이터를 활용하여 새소리 신호의 주파수 축을 인간의 청각 특성에 맞춘 멜 스케일(Mel scale)로 변환한 후, 이를 바탕으로 주파수 특성을 요약한 MFCC(Mel Frequency Cepstral Coefficients)를 추출하였다.

MFCC를 사용함으로써 다음과 같은 효과를 기대할 수 있다.

1. 새소리 데이터의 노이즈 영향 감소 및 명료성 향상

MFCC는 사람의 청각 메커니즘을 모사한 멜 스케일로 신호를 변환하기 때문에, 주변 환경 노이즈 및 불필요한 잡음으로부터 새소리 신호를 명확히 구분할 수 있다.

2. 복잡한 오디오 신호의 효과적인 패턴 추출

새소리와 같은 복잡하고 다양한 주파수 특성을 가진 오디오 신호에서도 특징을 효과적으로 요약하고 패턴을 명확하게 추출할 수 있다.

이러한 목적으로 MFCC를 활용하여 새소리 데이터를 효과적으로 처리하고, 궁극적으로 조류 종 분류 정확도의 향상과 모델 성능 개선을 기대하고자 한다.

3. 스펙트로그램 추출

기술 설명

스펙트로그램(Spectrogram)은 오디오 신호를 시간-주파수 영역으로 변환하여 시각적으로 표현한 것으로, 음성의 시간적 변화와 주파수 특성을 이미지 형태로 나타낸 것이다.

스펙트로그램은 음성 신호가 가지는 세부적인 주파수 변화 정보와 같은 풍부한 정보를 효과적으로 포함하여, 음성의 핵심 속성인 높이(pitch), 세기(intensity), 음색(timbre)을 충실히 표현한다.

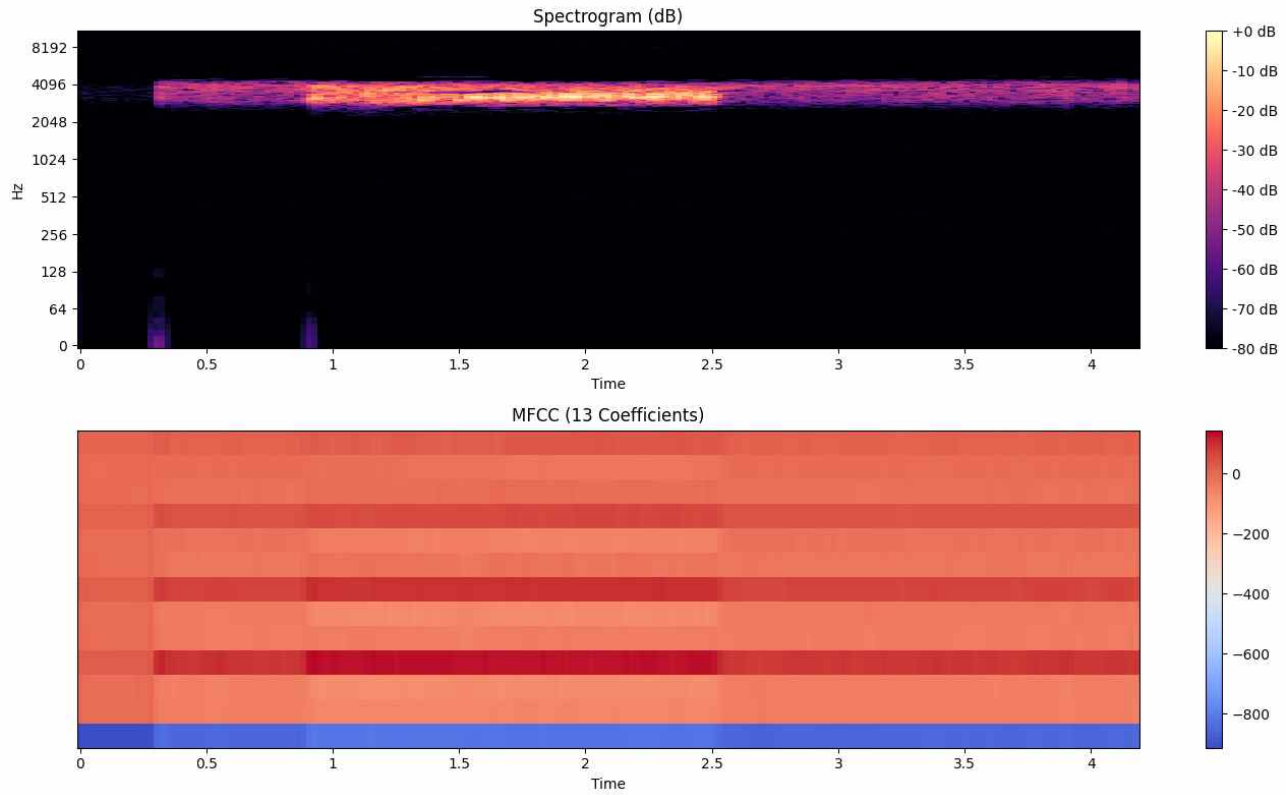
이번 프로젝트에서는 생성된 스펙트로그램 이미지를 일정한 크기(224×224)의 이미지 패치로 변환한 후, 이를 이미지 인식에 최적화된 딥러닝 모델의 입력으로 사용할 예정이다.

사용 목적

이번 딥러닝 모델의 최종 목적은 조류 종 분류 정확도를 극대화하는 것이므로, 오디오 신호의 세부적인 정보가 풍부한 스펙트로그램을 사용하여 정확도를 높이고자 하였다.

특히 스펙트로그램은 음성 신호의 세부 주파수 변화를 이미지 형태로 표현하여 ViT(Vision Transformer)와 EfficientNet 과 같은 이미지 기반 모델의 입력으로 적합하여, 이를 통해 각 조류 종별 미세한 음향적 특성을 효과적으로 포착하고 종 분류 정확도를 높이기 위해 스펙트로그램을 사용하였다.

적용 예시



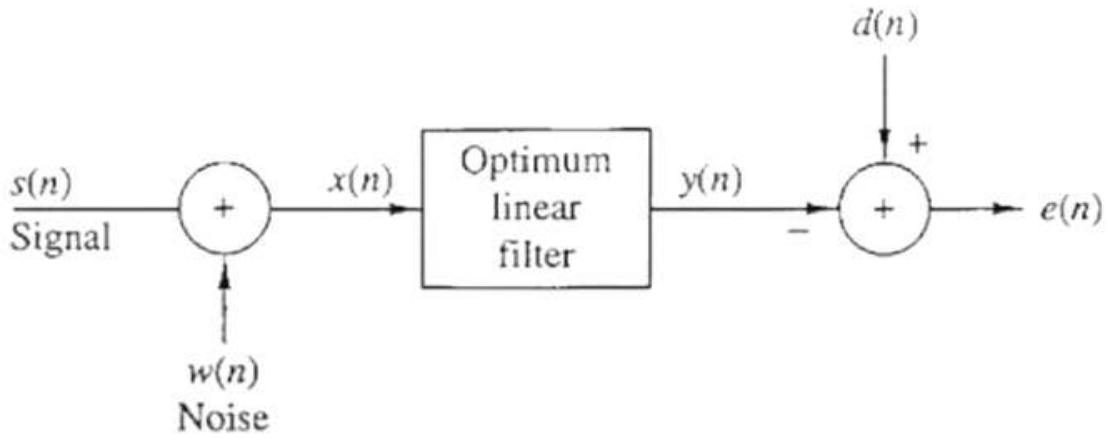
사용 기술

음성 분리 모델 Spleeter

오디오 파일 분석 중 녹음 이후 이에 관해 설명하는 형식의 오디오 파일이 상당수 존재한다는 것을 발견했다.
해당 형식의 경우 사람의 목소리가 절반 이상을 차지하기 때문에 학습에 방해되리라 판단했고, 이를 위해 음원에서 보컬과 반주를 분리해 내는 Spleeter 모델을 통해 사람의 목소리만 제거하는 방안을 생각했다.

이미지 파일을 학습시키는 방법 중 가장 유명한 방법은 CNN(합성곱 신경망) 모델이다. CNN은 입력으로 주어진 Tensor를 합성곱 필터를 통과시켜 특징을 추출하고 Pooling을 통해 규모를 축소해 연산량을 줄인다. 이 축소된 특징이나 패턴 정보를 학습해 분류해 내는 모델이

CNN이고, 분류에서 더 나아가 합성곱 및 Pooling 과정을 역추적해 해당 패턴을 생성하는 모델이 CNN에서 파생된 U-Net 모델이다. Spleeter의 경우에는 6층의 Encoder와 6층의 Decoder로 이루어진 12-layer U-Net 모델을 사용하며, 학습 데이터는 Deezer사의 음성분리 모델에 관한 논문에 사용된 내부 Dataset 을 사용한다. U-Net 모델을 통해 재생성된 스펙트로그램은 학습 데이터의 결과 음성 파일의 스펙트로그램과 비교해 최적화하는 과정을 갖게 되는데, 이때 손실함수는 L1 손실함수로 MAE를 사용하게 되고 최적화 알고리즘의 경우에는 모멘텀과 RMSprop 알고리즘을 섞은 Adam 최적화 알고리즘을 사용한다.



학습을 통해 완성된 모델은 사람의 목소리 스펙트로그램을 얻을 수 있다. 하지만 이 모델을 통해 원본에서 사람의 목소리가 아닌 사람의 목소리가 제거된 오디오 파일을 필요로 파일이 있어야 하므로 추가적인 과정을 거쳐야 한다. 원본 파일에서 특정 음성 신호를 추출한 뒤 해당 음성 신호의 역 파장을 더 해주면 될 것 같지만 스펙트로그램은 복소수이기 때문에 위상이 정확히 일치해야 단순 빼기가 가능하므로 Multi-channel Wiener Filtering이나 Soft masking을 사용해 사람의 음성 신호를 제거한다.

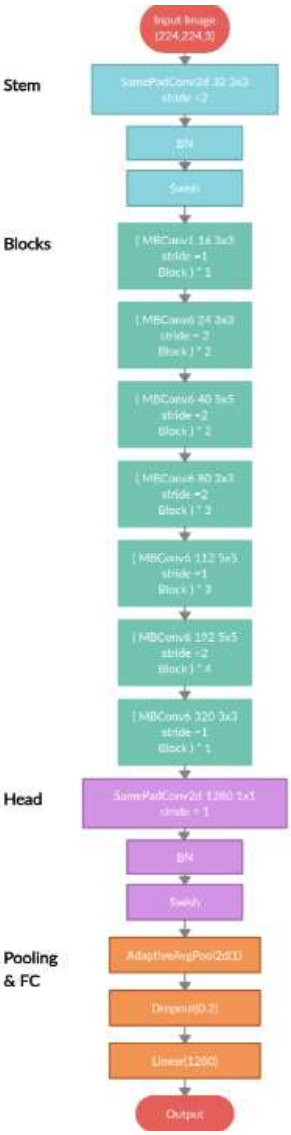
Wiener Filtering은 원본의 파워스펙트럼을 원본+추출의 파워스펙트럼으로 나눠서 사람 목소리가 이루는 주파수의 경우 원본에서 진폭을 줄여주고 아닌 부분은 통과시켜 주는 필터링 방식이다. 그리고 Soft masking의 경우 Winere filtering과 거의 유사한 식을 적용하기 때문에 일반화된 Wiener filtering이라고 할 수 있다.

Spleeter 모델의 목적은 보컬과 반주를 분리하는 것으로 음악 분야에서 사용되는 모델이다. 그러나 모델을 분석했을 때 해당 모델은 보컬, 즉 사람 목소리만 구분하도록 학습된 모델이고 사람 목소리의 제거는 추출한 목소리를 필터링 처리를 이용하기 때문에 자연물의 소리와 사람의 소리가 섞인 birdclef-2025의 학습 데이터에서 사람 목소리만 분리해 낸 자연물 소리를 얻을 수 있음을 확인했다.

EfficientNet

모델 소개

EfficientNet은 2019년 Google AI 팀에서 발표한 고성능 딥러닝 이미지 분류 모델로, 파라미터 수와 연산량을 최소화하면서도 정확도를 극대화한 것이 특징이다. 기존의 딥러닝 모델들이 단순히 깊이(depth)나 너비(width)를 증가시켜 성능을 향상시키려 했던 반면, EfficientNet은 이 세 가지 요소를 균형 있게 확장하는 Compound Scaling 기법으로 효율성을 상승 시킨다. 파라미터 수 대비 성능이 뛰어나기에 리소스가 제한된 BirdCLEF task에 적합할 것 같다 생각하여 적용시켜 보았다.



모델 구조

Table 1. **EfficientNet-B0 baseline network** – Each row describes a stage i with $\hat{L}_i$ layers, with input resolution $\langle \hat{H}_i, \hat{W}_i \rangle$ and output channels $\hat{C}_i$. Notations are adopted from equation 2.

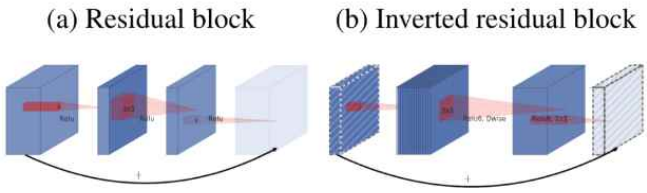
Stage i	Operator $\hat{\mathcal{F}}_i$	Resolution $\hat{H}_i \times \hat{W}_i$	#Channels $\hat{C}_i$	#Layers $\hat{L}_i$
1	Conv3x3	224×224	32	1
2	MBConv1, k3x3	112×112	16	1
3	MBConv6, k3x3	112×112	24	2
4	MBConv6, k5x5	56×56	40	2
5	MBConv6, k3x3	28×28	80	3
6	MBConv6, k5x5	14×14	112	3
7	MBConv6, k5x5	14×14	192	4
8	MBConv6, k3x3	7×7	320	1
9	Conv1x1 & Pooling & FC	7×7	1280	1

EfficientNet은 오른쪽 그림과 같은 구조로 이뤄져 있으며, 위 table 대로 구성되어 있다.

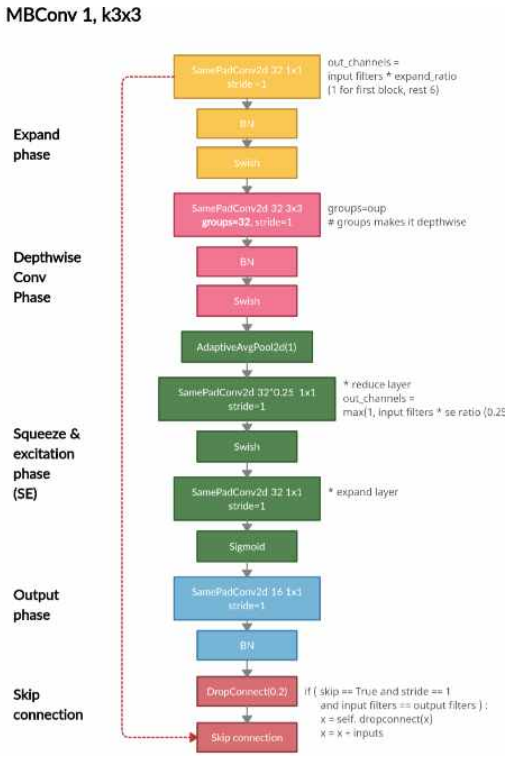
핵심 구조는 MBConv Block, Squeeze-and-Excitation(SE) 모듈, Swith Activation Function, Compound Scaling으로 이루어져 있다.

MBConv Block은 MobileNetV2 에서 도입된 Inverted Bottleneck 구조를 기만으로 하고, Depthwise Convolution과 Pointwise Convolution을 조합해 연산 효율을 높인다. 일반적인 CNN 블록은 넓은 채널에서 좁은 채널을 거쳐 넓은 채널로 향하는 구조를 가졌지만, MBConv는 그 반대로 좁은 채널에서 넓은 채널을 통해 좁은 채널로 향하는 구조를 사용한다.

MBConv Block의 구성 방식으로는 Pointwise Convolution (1x1 Conv)으로 채널 수를 늘려 다양한 특징을 뽑을 준비를 하고, Depthwise Convolution (3x3 Conv)으로 채널별로 독립적으로 공간적 특징(모양, 패턴 등)을 추출하여 연산량을 줄여준다. 마지막으로 다시 Pointwise Convolution (1x1 Conv)로 채널 수를 줄여 모델을 가볍게 유지한다.



위의 그림을 통해 MBConv Block은 Inverted residual block의 형태임을 알 수 있다.
오른쪽 그림은 실제 MBConv의 구조를 나타낸 것이다.



SE 모듈은 채널 간의 관계를 학습해 중요한 특징에 더 많은 가중치를 부여함으로써 표현력을 향상시킨다.

Swish Activation Function은 $f(x) = x \cdot \sigma(x)$의 식으로 나타내지며, 입력값에 0~1 사이 값을 곱해주는 부드러운 함수이다. ReLU보다 부드럽고 성능이 좋은 활성화 함수이기에 사용된다.

Compound SCaling은 너비, 깊이, 해상도를 하나의 스케일팩터로 통합하여 확장하고, 본 task에서 사용하는 EfficientNet-B0를 기준으로 스케일팩터를 증가시키며 모델을 확장해나가고 있다.
수하걱으로 단일 스케일링 팩터 ϕ를 정하고 다음과 같이 나눠서 수치를 키운다.
$depth = \alpha^{\phi}$, $width = \beta^{\phi}$, $resolution = \gamma^{\phi}$ (단, α, β, γ는 미리 정한 상수이며, $\alpha \times \beta \times \gamma \approx 2$가 되도록 조절해 계산량이 적당히 두 배씩 증가하게 한다. 따라서 기존 모델들에서 어느 한 쪽이 병목이 되는 등의 문제가 개선되고 계산량 대비 성능이 더 뛰어나다.

EfficientNet은 이와 같은 모델 구조를 가지고 있다.

BirdCLEF에서 적용

이번 task는 오디오 분류 과제이나 EfficientNet 모델은 이미지 분류 모델이기에 음성 데이터를 Spectrogram하여 이미지 형태로 변한한 뒤 이를 입력으로 사용하였다. 생성된 Mel Spectrogram은 2D 배열이며, 이를 0~255 범위의 RGB 이미지로 정규화해 변환했고, 사용할 모델인 EfficientNet-B0의 입력 크기에 맞게 244x244로 리사이즈한 뒤에 모델에 입력시켰다. 클래스 수에 맞춰 출력층을 수정하고, CrossEntropyLoss로 학습을 진행시켰으며, 사전 학습된 가중치를 사용하고, 일부 레이어는 고정시켜 전이학습 성능을 극대화했다.

ViT

ViT는 이미지 분류를 위해 고안된 모델이다.

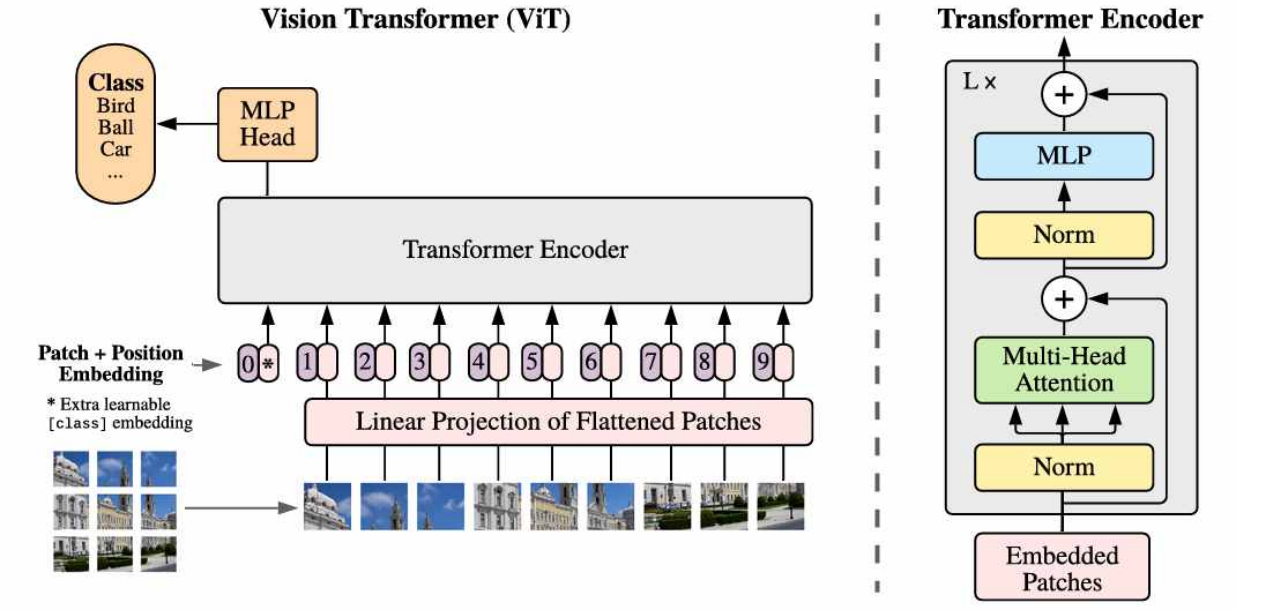
CNN 계열은 합성곱연산을 통해 지역적인 특성을 잘 포착하는 장점이 있지만, 전체 이미지를 포괄적으로 이해하는 데는 한계가 존재한다.

RNN에서도 장기 의존성이 있는 신호를 처리할 때 어려움이 있었으며, 이를 해결하기 위해 Attention 기반의 알고리즘이 도입되었다.

이러한 발전을 통해 등장한 구조가 바로 Transformer이다.

Transformer는 시계열 데이터를 포함한 다양한 입력에 대해 전체적인 맥락을 고려한 처리를 가능하게 하며, 비전(Vision) 분야로도 확장되어 ViT(Vision Transformer) 모델로 발전하였다.

ViT는 이미지 전체를 패치 단위로 분할하고, 각 패치를 선형 투영하여 순서를 유지한 채 Transformer에 입력함으로써 전체 이미지를 이해한다.



ViT는 본래 이미지 분류(Vision Task)를 위한 모델이지만, 본 프로젝트에서는 이 구조를 음성 분류에 적용하였다.

음성은 시계열 특성을 가지는 데이터이며, 이를 이미지처럼 처리하기 위해 스펙트로그램을 활용하였다.

입력은 스펙트로그램을 224×224 크기의 이미지 패치로 변환한 후, ViT 모델에 입력된다.

출력은 softmax함수를 통해 각 클래스에 해당할 확률을 나타내는 벡터로 도출된다.

클래스 간 데이터 불균형 문제를 고려하여, 정확도 대신 F1-score, Confusion Matrix를 주요 성능 평가 지표로 채택하였다.

참고 문헌

Spleeter
깃허브: https://github.com/deezer/spleeter

Laure Prétet 외 3인, "SINGING VOICE SEPARATION: A STUDY ON TRAINING DATA", in ICASSP 2019 – 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)

https://wowon-s.tistory.com/entry/%EC%98%81%EC%83%81%EC%B2%98%EB%A6%AC-%EC%98%81%EC%83%81%EC%9D%98-%EB%B3%B5%EC%9B%90%EA%B3%BC-%EC%9C%84%EB%84%88-%ED%95%84%ED%84%B0Wiener-filter?pidx=0

EfficientNet

https://lynnshin.tistory.com/53

https://arxiv.org/pdf/1905.11946

https://lynnshin.tistory.com/13

ViT

https://arxiv.org/abs/2010.11929