전이 학습과 SHAP 분석을 이용한 설명가능한 동물 울음소리 분류 기법

이재승, 문재욱, 박성우, 황인준
고려대학교 전기전자공학과
{jason2133, jaewookmo, psw5574, ehwang04}@korea.ac.kr

Explainable Animal Sound Classification Scheme using Transfer Learning and SHAP Analysis

Jaeseung Lee, Jaeuk Moon, Sungwoo Park, Eenjun Hwang
School of Electrical Engineering, Korea University

요 약

인간의 산업 활동으로 인하여 동물들의 생존이 위협받으면서, 동물의 서식 분포를 효과적으로 파악할 수 있는 자동 야생동물 모니터링 기술의 필요성이 점점 더 커지고 있다. 그중에서도 동물 소리 분류 기술은 시각적으로 식별이 어려운 동물에게도 효과적으로 적용할 수 있는 장점으로 인하여 널리 사용되고 있다. 최근 심층학습 기반의 분류 모델들이 좋은 판별 성능을 보여주고 있어 동물 소리 분류에 많이 사용되고 있지만, 희귀종과 같이 개체 수가 적어 데이터가 부족한 경우에는 학습이 제대로 이루어지지 않을 수 있다. 또한, 이러한 모델들은 모델 내부에서 일어나는 추론 과정을 알 수 없어 결과를 완전히 신뢰하고 사용하는 데 제약이 따른다. 이에 본 논문에서는 전이 학습을 통해 데이터 부족 문제를 고려하고, SHAP을 이용하여 분류 모델의 추론 과정을 해석하는 설명가능한 동물 소리 분류 기법을 제안한다. 실험 결과, 제안하는 기법은 지도 학습을 한 경우보다 분류 성능이 향상됨을 확인하였으며, SHAP 분석을 통해 모델의 분류 근거를 이해할 수 있었다.

1. 서론

지난 수십 년 동안, 인간의 산업 활동과 도시 개발은 동물 서식지를 파괴하고 파편화하는 등 야생동물에게 부정적인 영향을 미쳤다. 예를 들어, 산업 개발이 왕성했던 1970년 이후 전 세계의 야생동물 개체 수가 평균 68% 감소한 것으로 알려졌다[1]. 인간의 산업 활동에 대한 동물의 피해를 줄이기 위해서는, 다양한 동물들의 현황과 서식 분포 등을 지속해서 관찰하는 것이 중요하다. 과거에는 전문가가 직접 생태 현장을 조사하는 방법이 일반적이었지만, 이 방식은 많은 시간과 인력이 소모되는 단점이 있었다. 이에 대한 대안으로, 최근에는 음향 센서나 카메라를 활용한 자동 야생동물 모니터링 기술이 널리 사용되고 있다[2].

자동 야생동물 모니터링에서 효과적으로 활용되는 동물 소리 분류는 음향 센서를 이용하여 동물 울음소리를 수집하고, 동물의 종을 판별하는 기술이다. 이 방법은 주로 음향 데이터에 의존하기 때문에 동물의 크기가 작거나, 야행성이거나 위장색을 띄는 등 시각적으로 식별이 어려운 동물에게도 효과적으로 적용할 수 있다는 장점이 있다. 최근 심층학습 모델의 급속한 발전으로 인해, 다른 신호 처리 분야와 마찬가지로 동물 소리 분류에서도 합성곱 신경망과 같은 심층학습 모델을 통해 뛰어난 판별 성능을 달성하고 있다. 구체적으로, 동물 소리의 진폭만 고려하는 Waveform을 진폭과 주파수 모두 포함하는 2차원 이미지 데이터인 Spectrogram으로 변환하고, 이를 심층학습 모델의 입력 데이터로 활용하여 동물의 종을 분류한다. 이때, 심층학습 모델이 우수한 분류 성능을 달성하기 위해서는 목표하는 종에 대해 충분한 양의 데이터를 학습하는 것이 필요하다. 하지만, 동물 울음소리 데이터를 수집하는 과정에서 많은 양의 데이터를 확보하고 이에 대한 동물 종을 라벨링하는 데는 상당한 비용이 소모된다. 또한, 희귀종과 같이 개체 수가 적어 관찰이 어려운 동물의 경우 데이터 수집이 더욱 힘들다는 문제점이 있다. 이렇게 학습 데이터의 양이 충분하지 않을 경우, 심층학습 모델은 과적합 문제를 겪게 되어 일반화 능력이 떨어지고, 이로 인해 동물 소리 분류 성능이 크게 저하되는 문제가 발생한다.

이에 더해, 기존 심층학습 모델은 복잡한 연산으로 인해 모델의 판단에 대한 근거를 파악할 수 없어 흔히 '블랙박스'로 여겨진다. 이 때문에 모델이 입력 데이터의 어떤 특성을 기반으로 동물의 종을 분류하였는지 사용자가 알 수 없으며, 모델이 오판하더라도 원인을 알 수 없어 모델 개선을 어렵게 만든다. 그러므로, 모델이 추론한 결과에 대한 분석은 모델의 신뢰도와 정확성을 높이는 데 있어 필수적이다.

본 논문에서는 전이 학습[3]을 이용하여 데이터 부족 문제를 해결함으로써 동물 소리 분류 모델의 성능을 향상시키고, 설명 가능한 인공지능 기술인 SHapley Additive exPlanations (SHAP)[4]을 통해 모델의 분류 결과를 해석하는 기법을 제안한다. 먼저, 방대한 양의 이미지 데이터를 사전에 학습한 합성곱 신경망 기반 분류 모델의 가중치를 미세조정하여 전이 학습 기반의 동물 소리 분류 모델을 구축하였다. 이후에, 각 목표 종에 대한 분류 결과에 대하여 SHAP을 이용해 모델의 추론 과정을 분석하였다. 제안한 기법의 효과를 입증하기 위해 동물 소리 데이터만을 학습한 지도학습 모델과 제안하는 기법의 성능을 비교하였으며, 제안하는 기법은 분류 정확도 와 설명 가능성이 향상됨을 확인하였다.

본 논문의 구성은 다음과 같다. 2장에서는 제안하는 기법을 설명하고, 3장에서는 실험 환경 설정에 관하여 기술한다. 4장에서는 제안하는 기법의 정량적인 실험 결과와 모델 해석 결과를 보이고, 5장에서는 결론을 서술한다.

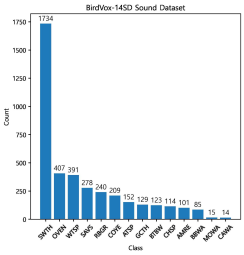
2. 제안하는 기법

2.1 데이터셋 구축 및 전처리

본 논문에서 제안하는 기법의 효과성을 검증하기 위하여, 공개적으로 사용 가능한 BirdVox-14SD[5] 조류 울음소리 데이터셋을 이용하였다. 본 데이터셋은 다양한 잡음을 포함한 야외 환경에서 수집된 것으로, 14개의 조류 종에 해당하는 음향 클립을 포함하고 있다. 그림 1은 본 데이터셋의 요약을 나타낸다. 본 논문에서는 음향 길이 0.5초로 구성된 총 3,992개의 조류 울음소리 데이터를 이용하였다.

동물 울음소리는 주파수, 지속시간, 속도와 같은 고유한 음향적 특성을 지니고 있다. 이러한 특성을 심층학습 모델이 효과적으로 학습하기 위해서는, 단순히 진폭만을 나타내는 1차원 데이터인 Waveform을 진폭과 주파수를 함께 고려하는 2차원 데이터인 Spectrogram으로 변환해야 한다. 본 논문에서는 각 동물 소리 파형에 대해 단기간 푸리에 변환 연산을 적용하여 다양한 주파수를 가지는 주기 함수들로 분해하였다. 이를 나타낸 Spectrogram이 고정된 크기를 갖도록 Zero-padding을 수행하였다. 그 결과, 모든 데이터는 128×128 크기의 1채널 Spectrogram으로 나타내었다. 그림 2는 BirdVox-14SD 데이터셋의 동물 소리 데이터 중 Waveform과 Spectrogram의 예시를 나타낸다.

(1) Class Distribution

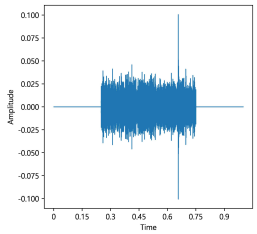


(2) Class Name

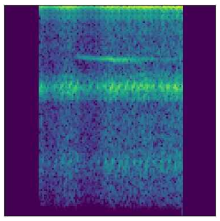
American Tree Sparrows (ATSP)
Chipping Sparrows (CHSP)
Savannah Sparrow (SAVS)
White-Throated Sparrow (WTSP)
Rose-Breasted Grosbeak (RBGR)
Gray-Cheeked Thrush (GCTH)
Swainson's Thrush (SWTH)
American Redstart (AMRE)
Bay-Breasted Warbler (BBWA)
Black-Throated Blue Warbler (BTBW)
Canada Warbler (CAWA)
Common Yellowthroat (COYE)
Mourning Warbler (MOWA)
Ovenbird (OVEN)

(그림 1) BirdVox-14SD 데이터셋 요약

(1) Waveform



(2) Spectrogram



(그림 2) 동물 소리 데이터 예시

2.2 전이 학습

전이 학습은 많은 양의 데이터를 사전에 학습한 모델의 가중치를 해결하고자 하는 작업에 맞게 미세조정하여 이용하는 훈련 기법이다[3]. 이 방법은 데이터가 부족하거나 특정 분야의 정보가 제한적인 상황에 유용하며, 기존에 학습된 지식을 활용하여 새로운 문제에 빠르게 대응할 수 있도록 한다. 본 논문에서는 대규모 이미지 데이터셋인 ImageNet을 사전에 학습한 합성곱 신경망 기반 모델을 이용하였다. 사전에 학습된 모델이 3채널 이미지를 사용하기 때문에, 1채널인 Spectrogram을 복사하여 3채널로 변환하였다. 이후, 사전에 학습된 모델로부터 전이된 가중치를 미세조정하여 동물 소리 분류 모델을 구축하였다. 이를 통해, 모델이 복잡한 동물 소리 특징을 효과적으로 학습할 수 있도록 하였다.

<표 1> 지도 학습과 제안 기법 간 동물 소리 분류 성능 비교

지표 / 모델명	Precision		Recall		F1 Score		Accuracy	
	지도 학습	제안 기법	지도 학습	제안 기법	지도 학습	제안 기법	지도 학습	제안 기법
ResNet	0.6455	0.8199	0.6728	0.7519	0.6549	0.7676	0.8723	0.8924
EfficientNet	0.6410	0.7283	0.6489	0.6868	0.6423	0.6964	0.8573	0.8924
MobileNet V3	0.5723	0.5733	0.6079	0.5498	0.5801	0.5519	0.8135	0.7985
ShuffleNet	0.5289	0.6463	0.5638	0.6223	0.5407	0.6315	0.7960	0.8486
RegNet	0.5546	0.7134	0.6134	0.7024	0.5675	0.6973	0.8073	0.8924
DenseNet	0.6346	**0.8700**	0.6550	**0.7561**	0.6356	**0.7871**	0.8573	**0.8999**

2.3 SHAP

SHAP은 이미지 내 각 픽셀의 SHAP value 계산을 통해 특정 영역이 최종 분류 결과에 어떻게 기여하는지를 분석하는 설명 가능한 인공지능 기법이다[4]. 이 값은 각 픽셀이 결과에 미친 영향의 정도를 수치적으로 나타내며, 이미지 분류에서는 이를 열 지도 형태로 시각화하여 모델이 중요하게 고려한 영역을 확인할 수 있다. 본 논문에서는 SHAP을 통해 동물 소리 분류 결과을 해석함으로써 분류 모델의 결정 과정을 파악하고자 하였다. 각 종을 판별하는 결과에 대해, 열 지도를 통해 표시된 각 픽셀의 색의 진함의 정도를 분석함으로써 해당 픽셀이 예측 결과에 얼마나 영향을 미쳤는지를 확인할 수 있다.

3. 실험 및 결과

3.1 실험 설정

본 논문에서는 제안하는 기법을 검증하기 위해 동물 소리 분류 실험을 수행하였다. 실험 데이터셋은 2.1에서 구축한 BirdVox-14SD 데이터셋을 사용하였으며, 전체 데이터셋을 8:2의 비율로 무작위로 나누어 훈련 및 평가 셋을 구축하였다. 정량적인 평가 지표로는 분류 성능을 측정하는 데 주로 사용되는 Precision, Recall, F1 Score, Accuracy를 선택하였다.

동물 소리 분류에 대한 제안 기법의 효과성을 검증하기 위해, 본 연구에서는 다양한 동물 소리 분류 모델들을 구축하였다. 최근 많은 연구들은 동물 소리 Waveform을 이미지 형태의 Spectrogram으로 변환하고 이를 입력으로 사용하는 합성곱 신경망 기반의 이미지 분류 모델들을 사용한다. 그러므로, 본 실험에서는 대표적인 이미지 분류 모델들인 ResNet[6], EfficientNet[7], MobileNet V3[8], ShuffleNet[9], RegNet[10], DenseNet[11]을 선택했으며, 각 모델마다 지도 학습과 전이 학습을 수행하여 비교 분석하였다. 이때, 지도 학습 기반의 모델들은 BirdVox-14SD 데이터셋의 훈련 셋을 사용하여 최대 50 Epoch 훈련했으며, 전이 학습 기반의 모델들은 ImageNet에서 사전 훈련한 뒤 앞서 기술한 훈련 셋에서 미세 조정하였다.

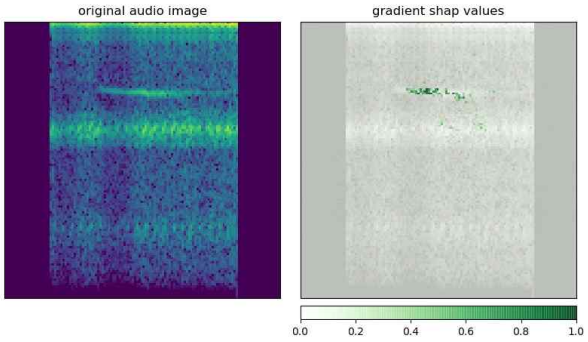
3.2 동물 소리 분류 모델 성능 평가

표 1은 지도 학습과 제안 기법 간 동물 소리 분류 모델들의 성능 평가 결과를 나타낸다. MobileNet V3를 제외한 모든 모델들은 전이 학습한 경우가 지도 학습한 경우보다 모든 평가 지표에서 더 높은 성능을 보였다. 특히, 전이 학습한 모델 가운데 DenseNet은 모든 평가 지표에서 가장 우수한 성능을 보였다. 이는 전이 학습 기반 모델이 이미 타 데이터셋에서 일반적인 특징을 식별하는 방법을 학습한 모델의 가중치를 이용함으로써 새로운 문제를 해결하기 위해 요구되는 데이터의 양을 상대적으로 줄였기 때문이라고 설명할 수 있다. 즉, 동물 소리 데이터와 같이 학습 데이터를 충분히 확보하기 어려운 경우에 대해서 전이 학습에 기반한 강화된 정보 표현 능력을 바탕으로 분류 모델의 성능을 향상시킬 수 있었다.

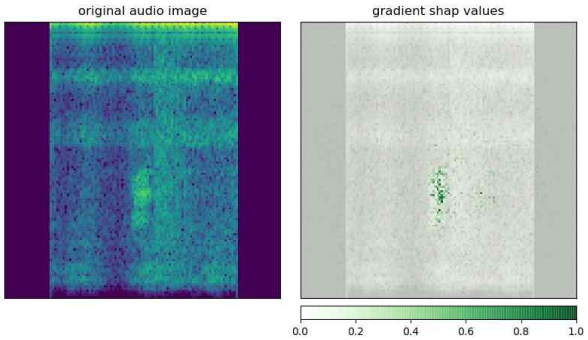
3.3 SHAP 분석을 이용한 판별 결과 해석

그림 3과 4는 SHAP을 이용하여 동물 소리 분류 모델이 각 목표 종을 판별하는 데 있어 중요한 특징이 무엇이었는지 시각화한 결과를 나타낸다. 여기서, 왼쪽 그림은 원본 Spectrogram을, 오른쪽 그림은 해당 Spectrogram에서의 SHAP 분석의 시각화 결과를 나타낸다. SHAP 분석 결과는 0-1 사이 값을 갖는 SHAP value를 초록색의 진한 정도로 표현하며, 색이 진할수록 스펙트로그램의 해당 픽셀이 모델의 판단에 긍정적인 영향을 주었음을 의미한다. 예를 들어, 그림 3의 가운데 상단 부분에 많은 초록색이 나타나 있는 것으로, 이 부분이 해당 Spectrogram을 Chipping Sparrow로 분류하는 데 중요한 역할을 했다는 사실을 알 수 있다. 이처럼, SHAP을 통해 Spectrogram에서 모델이 어느 부분을 중요하게 고

려하는지 시각적으로 확인함으로써 모델의 추론 결과에 대한 원인을 분석할 수 있었다. 이를 통해, 주어진 동물 소리에 대한 해당 모델의 판단 근거를 이해할 수 있었다.



(그림 3) Chipping Sparrow 분류 결과 해석



(그림 4) Mourning Warbler 분류 결과 해석

4. 결론

본 논문에서는 동물 소리 분류에서의 데이터 부족 문제를 해결하기 위해서 전이 학습을 이용하고, SHAP 분석을 통해 분류 결과에 대한 설명 가능성을 확보하는 기법을 제안하였다. 공개 조류 울음 소리 데이터셋을 이용한 동물 소리 분류 실험 결과, 데이터 부족 상황에서 전이 학습을 이용하는 것이 동물 소리 분류 모델의 성능을 개선한다는 점을 확인하였다. 또한, SHAP을 통해 모델이 분류 결과를 결정하는 데 있어 Spectrogram 내 어떠한 영역을 주로 주목하였는지 파악할 수 있었다.

향후 연구로, 동물 소리 분류에서의 데이터 부족 문제를 해결하기 위해 Model-Agnostic Meta Learning과 같이 최적화 기반의 메타 학습 모델을 이용하여 분류 성능을 높일 수 있는 기법을 연구할 계획이다.

사사문구

이 논문은 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.RS-2023-00262750, 인공지능 기반 소리인식기술을 이용한 감염병매개모기(일본뇌염 및 말라리아) 자동예찰 시스템 개발)

참고문헌

[1] R. Almond, M. Grooten, and T. Petersen. “Living Planet Report 2020 - Bending the curve of biodiversity loss”, in World Wide Fund for Nature, 2020.

[2] 이재승, 김은빈, 문재욱, 황인준. “앙상블을 사용한 Focal Loss 기반의 동물 소리 분류 기법”, 한국정보과학회 한국컴퓨터종합학술대회, 대한민국 제주, 2023, pp. 860-862.

[3] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. “A Comprehensive Survey on Transfer Learning”, *arXiv preprint*, arXiv:1911.02685, 2019.

[4] S. Lundberg and S. Lee. “A Unified Approach to Interpreting Model Predictions”, in NeurIPS, California, USA, 2017, pp. 4768-4777.

[5] A. Cramer, V. Lostanlen, A. Farnsworth, J. Salamon, and J. Bello. “Chirping up the Right Tree: Incorporating Biological Taxonomies into Deep Bioacoustic Classifiers”, in ICASSP, Barcelona, Spain, 2020, pp. 901-905.

[6] K. He, X. Zhang, S. Ren, and J. Sun. “Deep Residual Learning for Image Recognition”, in CVPR, Nevada, USA, 2016, pp. 770-778.

[7] M. Tan and Q. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”, in ICML, California, USA, 2019, pp. 6105-6114.

[8] A. Howard et al. “Searching for MobileNetV3”, in ICCV, Seoul, Korea, 2019, pp. 1314-1324.

[9] X. Zhang, X. Zhou, M. Lin, and J. Sun. “ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices”, in CVPR, Utah, USA, 2018, pp. 6848-6856.

[10] I. Radosavovic, R. Kosaraju, R. Girshick, K. He, and P. Dollar. “Designing Network Design Spaces”, in CVPR, Washington, USA, 2020, pp. 10428-10436.

[11] G. Huang, Z. Liu, L. Maaten, and K. Weinberger. “Densely Connected Convolutional Networks”, in CVPR, Hawaii, USA, 2017, pp. 4700-4708.