동물 소리 다중 분류를 위한 딥러닝 모델의 비교 분석

Comparative Analysis of Deep Learning Models for Multiclass Classification of Animal Sounds

전기전자공학부 2016170919 안현욱

지도교수 황인준

요 약

소리 데이터 분석 및 음성 인식은 전통적인 연구 분야 중 하나이다. 하지만 언어 인식의 연구는 활발히 진행되었던 한편, 환경 소리에 대한 분석은 언어와 달리 상대적으로 연구가 진행되지 않았다. 최근 반려 동물을 많이 키우며, 생태계 보전을 위해 동물 소리의 분류에 대한 필요성이 증가하고 있다. 이 연구는 ESC-50 소리 데이터 셋을 이용하여 이미지에서 활용되어 오던 더 고도화된 여러 모델을 활용하여 동물 소리를 다중 분류하고, 이를 비교하여 최적의 모델을 찾는다.

1. 서론

소리 인식 및 분류 연구는 컴퓨터 과학에서 전통적인 연구 분야 중 하나이다. 사람의 말을 인식하는 연구(Speech Recognition)의 경우, 1980년대부터 활발히 진행되었으며, 최근에는 인식의 정확도가 약 95%에 도달하였다[11]. 하지만 환경소리(Environment Sound)에 대한 분류(Classification)는 상대적으로 연구가 많이 진행되지 않았다. 그 한 예로, Google scholar 기준 2017년 이후 논문에 대한 키워드 검색 결과, “Speech Recognition”이 30만 여 건에 달하는 반면, “Environmental Sound Classification”은 5만여 건에 불과하다. 한편, 환경소리 분류 연구 중 동물 소리 분류는 최근 반려 동물 수의 증가와 환경 문제로 인해 수요가 계속적으로 증가하고 있다. 예를 들면, 1인 가구에서 반려 동물을 키우는 경우, 동물이 집에 혼자 남아있는 경우가 많았을 때 발생할 수 있는 사고를 예방하기 위해 CCTV, 카메라를 이용하여 반려 동물을 관찰하고 있으나, 카메라의 시야는 제한적이다. 동물 소리 분류는 이와 같은 상황에서 동물의 위치와 상황을 예상할 수 있는 기능을 제시할 수 있다.

최근에는 딥러닝이 유행하여 이미지의 인식과 분류에 대해 CNN(Convolutional Neuron Network) RNN(Recurrent Neural Network)과 같은 여러 딥러닝 모델이 활용되고 있다. 소리의 분류를 위해 CNN을 활용한 연구도[1] 진행되었으나, 이미지에서 활용되고 있는 고도화된 모델들을 활용한 연구는 아직 부족한 부분이 많다. 더불어, 새 소리 인식, 곤충 소리 인식과 같이 이진 분류(Binary Classification)에 대한 모델에 대한 제안은 많았으나[2-3], 다중 분류(Multiclass Classification)에 대한 제안은 적었다.

이 연구는 딥러닝을 활용하여 동물 소리를 인식하고 분류하기 위해 여러 모델에 따른 정확성을 비교해볼 것이다. 이미지에서 사용되어왔던 고도화된 모델들을 활용하여 소리 데이터 셋 ESC-50 Dataset[4]을 학습시켜 성능과 정확도를 비교할 예정이다. 2장에서는 관련 연구, 3장에서는 활용된 소리 데이터셋과 딥러닝 네트워크에 대한 소개를 할 것이며, 4장에서는 각각의 네트워크에 대한 학습 과정과 결과, 5장에서는 결론을 서술한다.

2. 관련 연구

새 소리에 대한 데이터 셋 및 연구는 분류가 상대적으로 쉬워 이에 대한 분석 및 분류는 2015년부터 진행되었다.[5] 특히 최근 논문에서는 환경 모니터링 및 생태계 보존을 위하여 여러 동물 소리를 한 번에 분석하여 섞여 있는 소음에 대해 동물들을 여러 클래스로 분류할 수 있는 모델을 만들었다[6].

3. 데이터셋과 모델

1) ESC-50 데이터 셋

ESC-50 데이터셋은 freesound.org에서 2000개의 환경 소리 녹음을 5개의 대분류(Animals, Natural soundscapes & water sounds, Human_non-speech sounds, interior/domestic sounds, Exterior/urban noises)로 분류 되며, 각각의 대분류는 10개의 소분류로 다시 나누어지며, 각 소리파일은 5초 길이로, 44.1kHz, mono채널인 .wav파일로 구성되어있다.

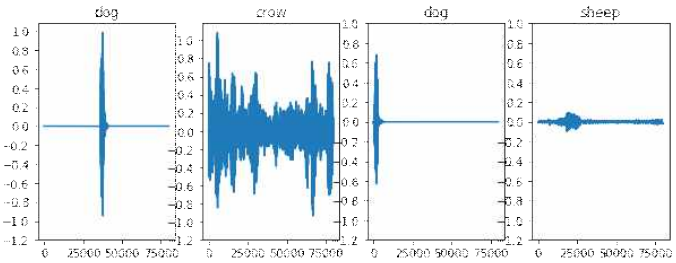

Fig.1 Waveform of sound of animals

이 연구에서는 위 데이터 셋에서 Animals 내에 있는 소분류 10개(Dog, Rooster, Pig, Cow, Frog, Cat, Hen, Insects, Sheep, Crow)를 분류할 예정이다. 각각의 소분류는 40개의 소리 파일로 구성되어있으며, 데이터셋을 증강하기 위해 원본 소리에 변조를 가하였다. 늘림(Stretching), 역재생(Reverse), 역 위상(Phase Reverse), 백색 소음 첨가(Adding White Noise), 순서 변경(Shifting)과 같이 원본 소리에 변조를 첨가하여 확장된 ESC-50 데이터 셋(Extended ESC-50 Dataset)을 구성하였다.

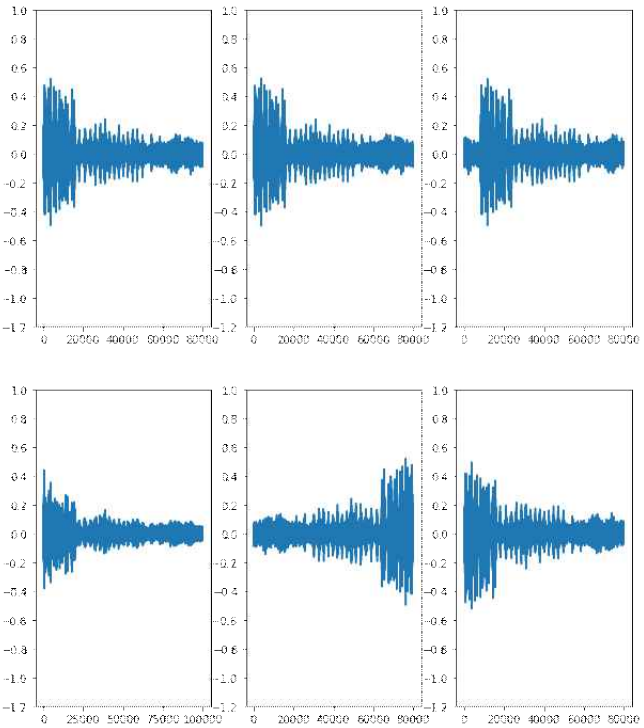

Fig.2 Waveforms of Extended dataset, from left-topmost, original, adding white noise, shifting, stretching, reverse, phase reverse

2) 데이터 전처리

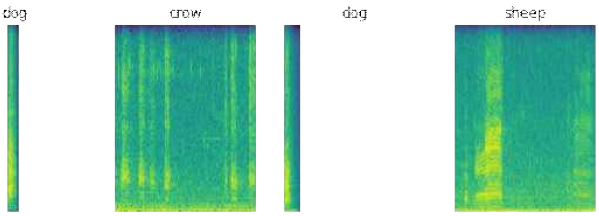

Fig.3 Spectrogram of animal sound

CNN은 2차원 정보의 보존으로 학습의 성능이 더 뛰어나다. 이를 활용하기 위해 시간에 대해 1차원인 소리 데이터를 일정한 구간에 대해 짧은 시간마다 각각 STFT(Short-Time-Fourier-Transform)하여 차원을 증강시켰다. 짧은 시간에 대한 STFT의 결과는 주파수 성분을 나타내며, 일정한 구간에 대해 짧은 시간마다 STFT를 하게 되면 각각의 짧은 시간마다의 주파수 성분을 얻어 해당 구간의 주파수 변화를 알 수 있다.

구체적으로, 25ms의 짧은 시간을 STFT하여 10ms간격으로 hop하여 소리 데이터의 스펙트럼(spectrogram)을 구성한다. 멜 스펙트럼(Mel Spectrogram)은 데이터를 양자화시키기 위해 64bit로 스펙트럼을 맵핑(Mapping)하여, 7500Hz까지의 주파수를 분석할 수 있도록 한다. 그 후, log를 취하여 로그 멜 스펙트럼(Log Mel Spectrogram)을 구한다. 이를 0.96s 만큼의 구간을 통하여 96프레임 수와 각 프레임의 64bit의 값을 (96,64)의 2d 형태로 구성한다. 일정한 구간을 50% overlapping (0.48s) 하여 0.96s마다 소리를 분류할 수 있도록 만한다.

3) 모델

모델은 다음과 같은 모델을 이용하였다.

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

Fig.4 EfficientNetB0

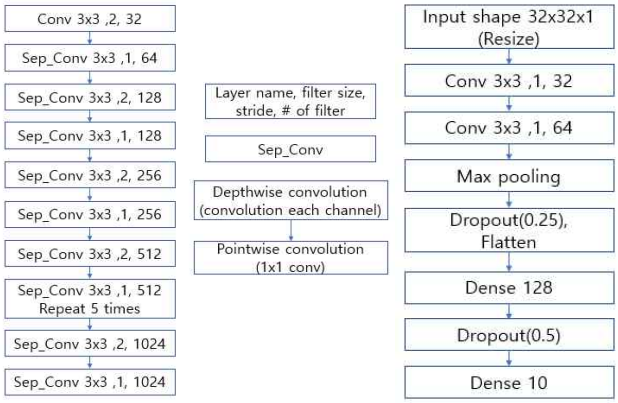

Fig.5 Yamnet

Fig.6 CNN

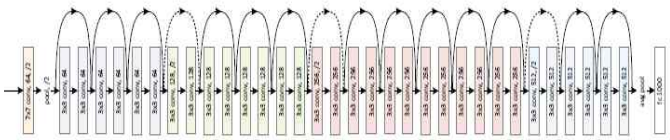

Fig.7 ResNet

-Traditional CNN(Convolutional Neuron Network)

이미지에서 사용되었던 전통적인 딥러닝 알고리즘으로, 이 연구에서는 Fig.6과 같은 CNN을 이용한다. Dropout을 이용하여 과적합(Overfitting)을 방지한다.

-Yamnet[7]

Yamnet은 Mobilenet-v1[8]을 기반으로 만들어진 모델이며, Google의 Audioset 데이터베이스를 이용하여 미리 학습된 Weight로 521가지의 클래스를 구분할 수 있도록 구성되었으며, embedding을 통해 분류 단계 전 1024단계로 출력이 가능하다. 이 연구에서는 Yamnet을 통과한 후 embedding으로 출력한 값들을 Full-linked NN을 이용하여 10개의 클래스를 분류할 수 있는 모델로 이용할 것이다.

Mobilenet-v1은 depthwise seperable convolution을 이용하여 파라미터의 수를 줄이고, 체널 간의 정보를 공유할 수 있는 장점이 있다. 이를 통해 빠른 학습과 추론이 가능하다는 장점이 있다. Yamnet의 구조는 Fig.5와 같다.

-ResNet[9]

ResNet는 많은 layer를 이용한 모델로, 네트워크가 깊어지면 gradient가 사라지는 현상으로 인해 오히려 학습이 제대로 되지 않는 것을 해결하기 위해 만들어진 네트워크이다. Convolution layers로 구성된 블록이 5개가 있으며, Convolution을 거친 값과 거치지 않은 값을 서로 더하여 이전의 값이 후반 층에 더 큰 영향을 미칠 수 있게 된다. ResNet의 구조는 Fig.7과 같다.

-EfficientNet[10]

EfficientNet은 CNN의 정확도를 최대로 높이기 위해 고안된 모델이다. 고도화된 모델들은 정확도를 높이기 위해서 layer의 깊이를 깊게 하거나 (e.g. ResNet), 필터의 개수를 늘린다(e.g. Mobilenet). EfficientNet은 NAS기법을 이용하여 가장 효율적인 CNN 모델인 EfficientNet-B0를 구성하며, 구조는 Fig.4와 같다. 구성된 모델에 대해 depth, width, resolution의 scaling factor를 구해 가장 효율적인 값을 이용하여 EfficientNet-B1부터 B7까지 구성하게 된다.

이 연구에서는 CNN을 제외한 네트워크는 10가지의 클래스로 분류할 수 있도록 네트워크 후반부에 Full-Linked NN을 구성하였다.

4. 실험 및 결과

실험 환경은 Intel i7-6700 CPU @ 3.40GHz, NVIDIA GeForce GTX 970, 16GB RAM, Windows 10 64bit OS에서 실행 되었다. 학습을 위한 초매개변수(Hyper parameter)는 Learning rate 1e-4, 최적화 알고리즘은 Adam으로, loss는 cross-entropy를 이용하고, epoch는 100으로 설정하였다.

학습을 위해 ESC-50 데이터셋에서 각 클래스별로 40개의 음성파일(총 400개) 중 60%을 트레이닝 셋으로, 20%을 검증 셋으로, 20%을 테스트 셋으로 이용하였다. 예측(Prediction)은 freesound.org의 Dog Barking .wav 파일을 이용하여 테스트한 결과이다.

Yamnet은 68.3%의 정확도를 보여주었으며, CNN은

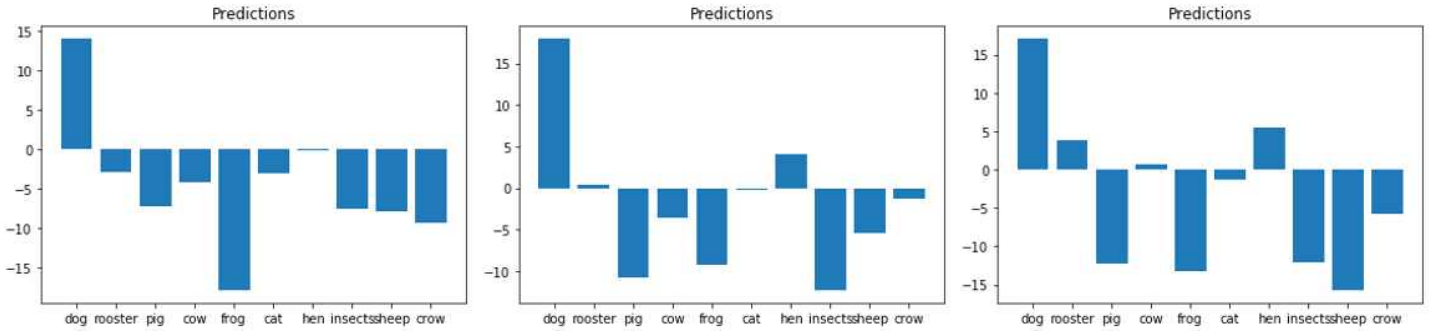

Fig.8 Prediction of dog sound, left yamnet, middle ResNet50V2, right EfficientNetB0

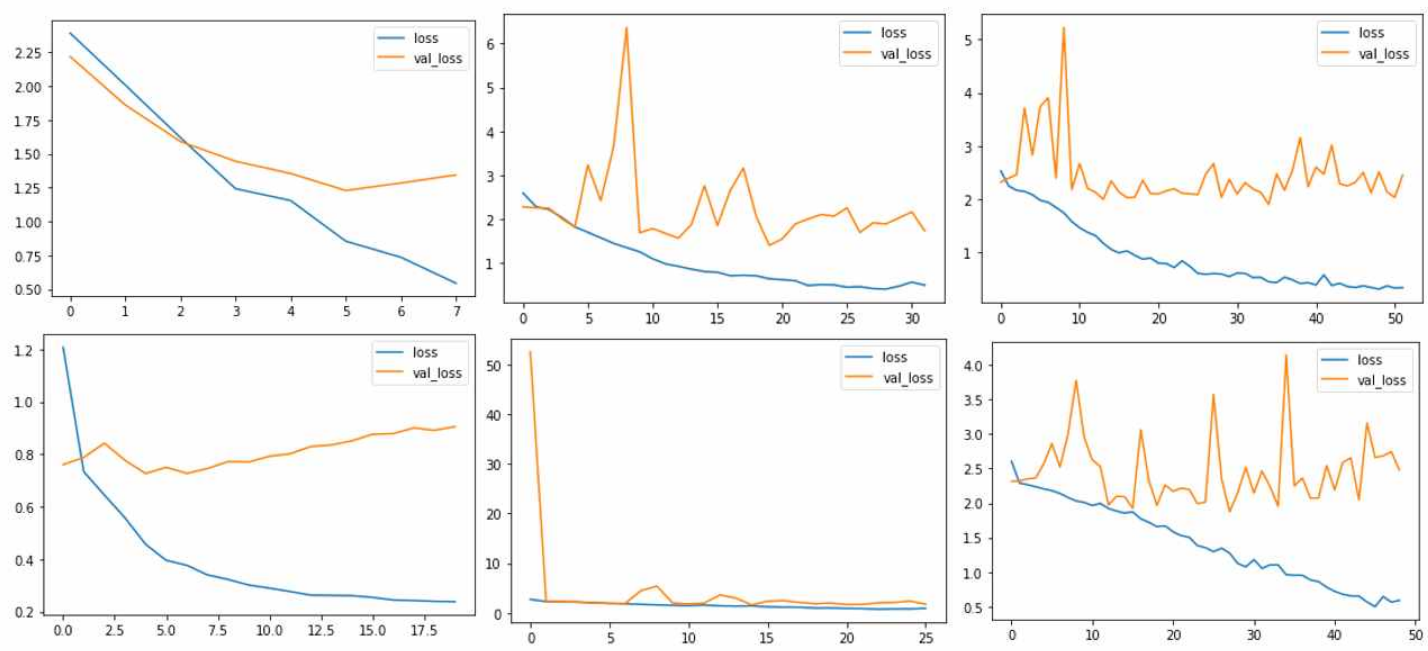

Fig.9 Training Result, left-top traditional CNN, left-bottom yamnet, middle-top ResNet50V2, middle-bottom ResNet152V2, right-top EfficientNetB0, right-bottom EfficientNetB3

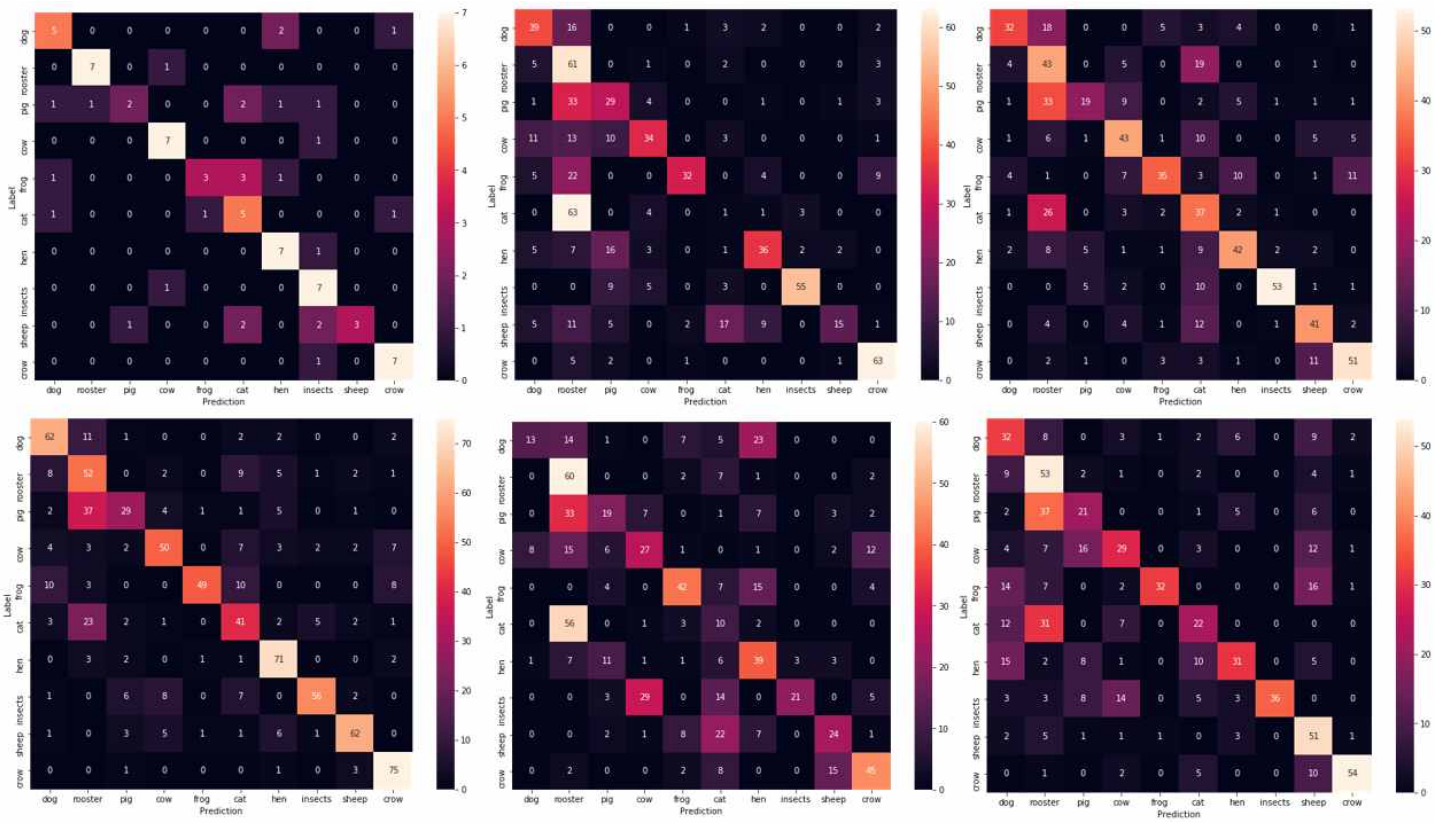

Fig.10 Confusion matrix, left-top traditional CNN, left-bottom yamnet, middle-top ResNet50V2, middle-bottom ResNet152V2, right-top EfficientNetB0, right-bottom EfficientNetB3

66%의 정확도를, EfficientNetB0은 55.8%의 정확도를, EfficientNetB3은 51.4%의 정확도를, ResNet50V2는 50.8%의 정확도를, ResNet152v2는 42.4%의 정확도를 보여주었다. Fig.9은 학습에 따른 Loss의 변화이며, Fig.10는 각각의 네트워크에 대한 Confusion Matrix이다. Fig.8은 학습된 모델을 이용하여 예측한 결과이다.

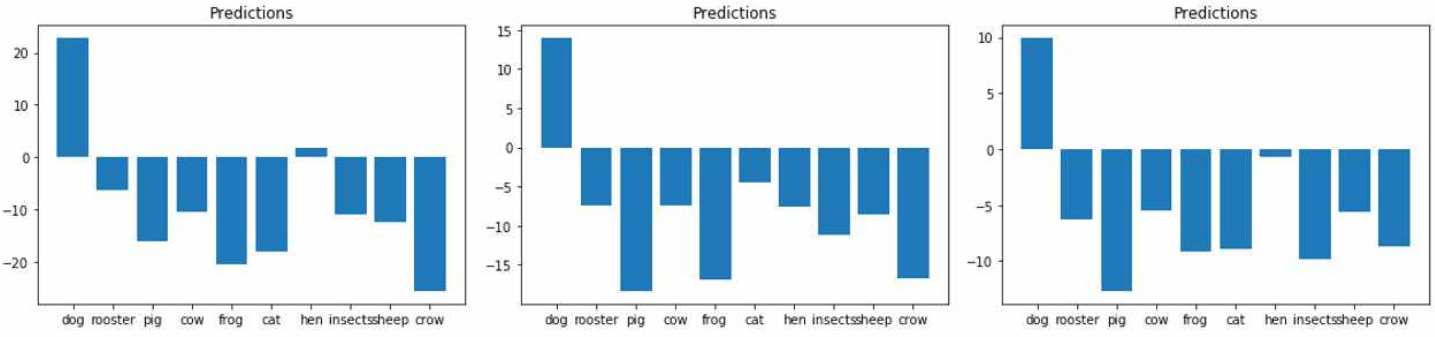

Fig.11 Prediction of dog sound trained by extended dataset, left yamnet, middle ResNet50V2, right EfficientNetB0

데이터를 증강하여 각 클래스별로 240개의 음성파일(총 2400개) 중 60%을 트레이닝 셋으로, 20%을 검증 셋으로, 20%을 테스트 셋으로 이용하여 각 모델에 대해 적용시킨 경우, yamnet은 66%의 정확도를, EfficientNetB0은 56.7%의 정확도를, ResNet50V2은 58.4%의 정확도를 보여주었다. Fig.12는 각각의 네트워크에 대한 Confusion Matrix이며, Fig.11은 학습된 모델을 이용하여 예측한 결과이다.

Table.1은 각 모델의 정확도를 나타낸 것이다.

Model		Accuracy (%)	Accuracy(%) (Extended Dataset)
CNN		66.3	.
Yamnet		68.3	66
ResNet	50V2	50.8	58.4
	152V2	42.4	.
Efficient Net	B0	55.8	56.7
	B3	51.4	.

Table 1. Accuracy of Models

5. 결론

Yamnet은 Audioset을 이용하여 pre-trained되었기 때문에 정확도가 가장 높게 나온 것으로 보인다. ESC-50의 데이터셋 크기가 크지 않아 학습가능한 데이터가 적기 때문에, Yamnet에 비해 높은 정확성을 보이지 못했다. 하지만 그럼에도 불구하고 EfficientNet은 약 55%의 정확성을 보여주었으며, 더 많은 데이터를 이용하여 학습을 하게 되면 Yamnet보다 더 높은 정확성을 보일 수 있다고 예상한다.

데이터 셋이 증강된 경우 EfficientNet과 ResNet이 모두 더 높은 정확성을 보여주었으며, 이는 데이터셋의 크기가 커지고 학습 데이터가 많아진다면 더 높은 정확성을 보일 것임을 알 수 있다.

이미 학습된 yamnet의 경우 증강된 데이터셋을 이용하여 학습을 하였을 때 정확성이 오히려 약간 감소하는 것을 볼 수 있었으며 이는 큰 데이터 셋을 이용하여 pre-trained된 모델에 대해서 증강된 데이터 셋을 이용한 학습은 효율적이지 않다는 걸 알 수 있다.

모델의 깊이(depth)와 너비(width)는 모델의 정확성을 높이지만, 어느 정도 이상으로 커지게 된다면 오히려 정확성이 떨어지는 것을 알 수 있다. 입력의 크기에 따라 적절한 깊이의 모델을 선정하는 것이 중요함을 알 수 있다.

여러 모델을 이용하여 동물 소리를 분류함으로써, Yamnet과 EfficientNet이 높은 정확도를 가져 이 네트워크에 대해 더 많은 데이터를 이용하여 학습한다면 더 높은 정확성으로 동물 소리 분류를 할 수 있을 것이다. 동물 소리 분류를 통해 생태계 유지 및 야생 동물 보호를 위한 동물 분류에 사용될 수 있으며, 후에 동물 소리 탐

Fig.12 Confusion Matrix trained by extended dataset, left yamnet, middle ResNet50V2, rigth EfficientNetB0

지를 통해 동물 위치 추정 및 추적에 이용될 수 있으며, 이를 통해 반려동물 관리를 더 용이하도록 할 수 있다.

참고문헌

[1] Piczak, Karol J. “Environmental sound classification with convolutional neural networks.“ 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP). IEEE, 2015.

[2] F. Briggs, Y. Huang, R. Raich, K. Eftaxias, Z. Lei, W. Cukierski, S. F. Hadley, A. Hadley, M. Betts, and X. Z. Fern, “The 9th annual MLSP competition: new methods for acoustic classification of multiple simultaneous bird species in a noisy environment,“ in Proc. of 2013 IEEE international workshop on machine learning for signal processing (MLSP), pp. 1-8, 2013.

[3] D. Pimentel, and M. Burgess, “Environmental and economic costs of the application of pesticides primarily in the United States,“ Integrated pest management, pp. 47-71, 2014.

[4] Piczak, Karol J. “ESC: Dataset for environmental sound classification.“ Proceedings of the 23rd ACM international conference on Multimedia. 2015.

[5] D. Stowell, M. Wood, Y. Stylianou, and H. Glotin, “Bird detection in audio: a survey and a challenge,“ in Proc. of 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), pp. 1-6, 2016

[6] Kim, Chung-Il, et al. “Animal Sounds Classification Scheme Based on Multi-Feature Network with Mixed Datasets.“ KSII Transactions on Internet and Information Systems (TIIS) 14.8 (2020): 3384-3398.

[7] Transfer Learning with YAMNet for environmental sound classification https://www.tensorflow.org/tutorials/audio/transfer_learning_audio

[8] Howard, Andrew G., et al. “Mobilenets: Efficient convolutional neural networks for mobile vision applications.“ arXiv preprint arXiv:1704.04861 (2017).

[9] He, Kaiming, et al. “Deep residual learning for image recognition.“ Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] Tan, Mingxing, and Quoc Le. “Efficientnet: Rethinking model scaling for convolutional neural networks.“ International Conference on Machine Learning. PMLR, 2019.

[11] SUMMA LINGUAE-Language Technology-A Complete Guide to Speech Recognition Technology