한국음향학회지 제39권 제5호 pp. 406~413 (2020)
The Journal of the Acoustical Society of Korea Vol.39, No.5 (2020)
https://doi.org/10.7776/ASK.2020.39.5.406

pISSN : 1225-4428
eISSN : 2287-3775

시간 축 주의집중 기반 동물 울음소리 분류

Temporal attention based animal sound classification

김정민,[1] 이영로,[1] 김동현,[1] 고한석[1†]

(Jungmin Kim,[1] Younglo Lee,[1] Donghyeon Kim,[1] and Hanseok Ko[1†])

[1]고려대학교 전기전자공학과

(Received July 22, 2020; accepted August 24, 2020)

초 록: 본 논문에서는 조류와 양서류 울음소리의 구별 정확도를 높이기 위해 게이트 선형유닛과 자가주의 집중 모듈을 활용해서 데이터의 중요한 부분을 중심으로 특징 추출 및 데이터 프레임의 중요도를 판별해 구별 정확도를 높인다. 이를 위해 먼저 1차원의 음향 데이터를 로그 멜 스펙트럼으로 변환한다. 로그 멜 스펙트럼에서 배경잡음같이 중요하지 않은 정보는 게이트 선형유닛을 거쳐 제거한다. 그러고 난 뒤 시간 축에 자가주의집중기법을 적용해 구별 정확도를 높인다. 사용한 데이터는 자연환경에서 멸종위기종을 포함한 조류 6종의 울음소리와 양서류 8종의 울음소리로 구성했다. 그 결과, 게이트 선형유닛 알고리즘과 시간 축에서 자가주의집중을 적용한 구조의 평균 정확도는 조류를 구분했을 때 91 %, 양서류를 구분했을 때 93 %의 분류율을 보였다. 또한, 기존 알고리즘보다 약 6 % ~ 7 % 향상된 정확도를 보이는 것을 확인했다.

핵심용어: 음향이벤트 인식, 합성 곱 신경망, 자가주의집중, 게이트 선형유닛

ABSTRACT: In this paper, to improve the classification accuracy of bird and amphibian acoustic sound, we utilize GLU (Gated Linear Unit) and Self-attention that encourages the network to extract important features from data and discriminate relevant important frames from all the input sequences for further performance improvement. To utilize acoustic data, we convert 1-D acoustic data to a log-Mel spectrogram. Subsequently, undesirable component such as background noise in the log-Mel spectrogram is reduced by GLU. Then, we employ the proposed temporal self-attention to improve classification accuracy. The data consist of 6-species of birds, 8-species of amphibians including endangered species in the natural environment. As a result, our proposed method is shown to achieve an accuracy of 91 % with bird data and 93 % with amphibian data. Overall, an improvement of about 6 % ~ 7 % accuracy in performance is achieved compared to the existing algorithms.

Keywords: Audio event classification, Convolution Neural Network (CNN), Self-attention, Gated Linear Unit (GLU)

PACS numbers: 43.60.Bf, 43.60.Uv

I. 서 론

환경영향평가는 특정 사업이 환경에 영향을 미치게 될 요인들을 제거하거나 최소화하기 위해 사전에 해당 지역의 환경 영향을 분석하여 검토하는 과정을 말한다. 이 과정에서 멸종위기종 보호 및 자연생태계를 보전하기 위해 해당 지역의 서식 종뿐만 아니라 종의 개체 수를 파악하는 일은 매우 중요하다. 하지만 평가를 할 수 있는 전문 인력이 부족하고 조사를 할 때 오랜 시간이 필요하므로 평가 동안 적절한 조사를 수행하기 어렵다. 이를 위해 전문가를 대신해 오디오 신호를 이용한 종 식별 및 개체 수를 파악하는 연구가 진행되고 있다.[1]

†**Corresponding author:** Hanseok Ko (hsko@korea.ac.kr)
Department of Electronics and Computer Engineering, Engineering Building Room 419, Korea University Anam Campus, 145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
(Tel: 82-2-3290-3239, Fax: 82-2-3291-2450)

기존 연구에서 동물 울음소리 인식을 위해 1차원의 오디오 데이터를 log-Mel spectrogram과 Mel Frequency Cepstral Coefficients(MFCC) 같이 시간과 주파수를 동시에 표현할 수 있는 2차원 데이터로 변환한다. 이후 변환된 데이터를 Convolution Neural Network (CNN) 같은 신경망 구조를 통해 소리의 특징을 추출한 뒤 Support Vector Machine(SVM), CNN 혹은 Fully-connected Neural Network(FNN) 등의 분류기를 활용해 종을 분류한다.[2-5]

하지만 기존의 연구는 동물 울음소리를 녹음할 때 같이 녹음되는 배경잡음에 관한 연구는 미흡하다. 보통 표적이 명확하고 데이터 대부분을 차지하고 있는 이미지를 분류하는 것과는 달리 동물 울음소리 같은 음향 데이터는 정보가 이미지보다 불명확하고 원하는 정보의 길이가 일정하지 않은 등 비정상성 특징을 가지며, 데이터 녹음 과정에서 발생하는 배경잡음을 항상 포함하고 있다. 이러한 배경잡음은 분류기 학습 과정에서 바이어스로 적용할 수 있으므로, 배경잡음에 강인한 음향이벤트 특징 추출이 중요하다. 또한, 단발적으로 발생하는 음향이벤트의 경우 프레임별 레이블이 존재하지 않기 때문에 학습 과정에서 문제가 발생할 수 있다.

최근 이런 문제를 해결하기 위해 일부 음향이벤트 분류 구조에서는 배경잡음에 강인한 특징을 추출할 수 있는 게이트 선형유닛 Gated Linear Unit(GLU)[6,7]과 별도의 사전정보 없이 프레임별 중요도 판별을 통해 음향이벤트 인식 정확도를 높일 수 있는 자가주의집중[8] 알고리즘을 적용하고 있다.

이를 토대로 본 논문에서는 자연환경에서 직접 녹음한 동물울음 소리를 2차원 데이터로 변환된 음향 데이터에서 프레임별 중요도를 판별하고 중요하지 않은 정보는 최대한 제거해 분류 정확도를 높이는 자가주의집중기법을 제안한다. 자가주의집중기법을 사용하면 프레임별 중요도를 판별할 수 있지만, 프레임 안에 배경잡음을 제거할 수 없으므로 GLU를 사용해 전체 입력데이터에서 잡음을 최대한 없앨 수 있는 특징을 활용한 구조를 제안한다. 제안한 구조를 통해 자연환경에서 울음소리를 녹음할 때 함께 포함된 바람이나 빗소리 등 불필요한 배경잡음을 최대한 제거하고 중요한 부분에 집중하여 분류 정확도를 높일 수 있다.

본 논문의 구성은 다음과 같다. 2장은 본 논문에서 제안하는 방법, 3장은 실험 및 결과, 4장은 토의, 5장은 결론으로 구성하였다.

II. 제안하는 방법

본 논문에서 사용한 데이터는 정확한 레이블 값이 있지만, 레이블이 데이터의 어떤 부분에 해당하는지 나타나 있지 않은 약한 레이블 데이터이다. 따라서 GLU를 통해 중요 정보에 대한 값은 살리고 배경잡음과 같이 불필요한 정보는 제거한다. 그리고 GLU를 통해 나온 데이터는 시간 축에서의 자가주의집중 기법으로 별도에 사전정보 없이 프레임별 중요도 판별해 정확도를 높인다. 이를 토대로 CNN과 FNN을 이용한 구조와 본 논문에서 사용한 GLU 그리고 자가주의집중 기법을 소개한다.

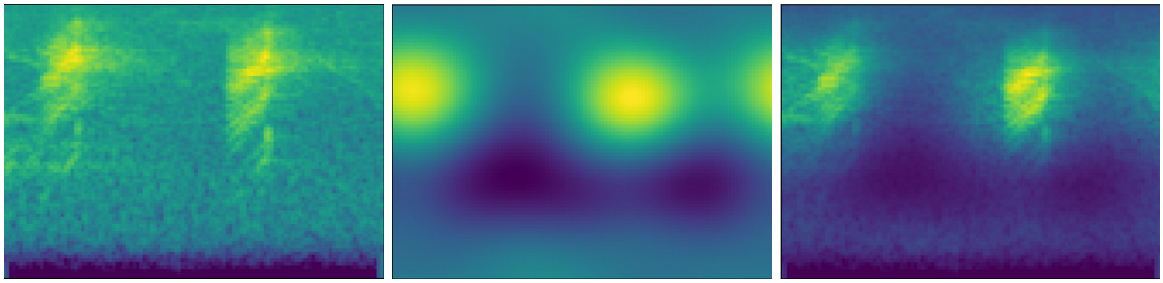
2.1 게이트 선형유닛(GLU : Gated Linear Unit)

동물 울음소리분류에서 데이터의 중요한 부분을 중심으로 특징을 추출하기 위해 게이트 선형유닛을 도입한다. GLU를 사용함으로써 배경잡음같이 불필요한 정보를 제거하면서 구분하고자 하는 데이터를 집중시키는 효과가 있다. GLU는 서로 다른 2개의 합성 곱 층을 활용한다. Eq. (1)처럼 하나의 합성 곱에서 나온 특징과 다른 합성 곱에서 나온 특징에 활성화 함수인 Sigmoid 함수를 통과한 값을 곱해 입력값의 정보를 제어한다.

$$h(x) = (X * W_1 + b_1) \otimes \sigma(X * W_2 + b_2), \qquad (1)$$

여기서 X는 입력데이터 W_1, b_1, W_2 그리고 b_2는 합성 곱 신경망으로 훈련할 파라미터다. σ는 sigmoid 함수, 그리고 $\otimes$는 요소별 곱셈이다. Sigmoid 함수를 통과한 특징을 게이트 값이라고 하는데, 이 값이 1에 가깝다면 해당 부분의 정보는 유지하고 반대로 0에 가깝다면 그 정보는 무시하고 다음 층의 입력으로 사용한다.[9]

Fig. 1(a)는 실제 동물 울음소리의 입력데이터를



(a) Input data before apply GLU (b) GLU feature map of input data (c) Output data after GLU layer

Fig. 1. (Color available online) Diagram when GLU is applied to data.

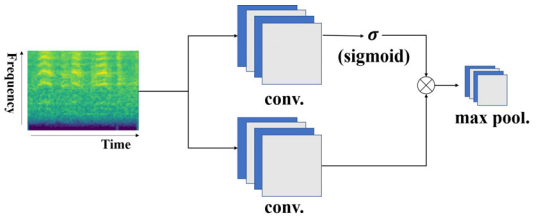


Fig. 2. (Color available online) Diagram of a GLU-block.

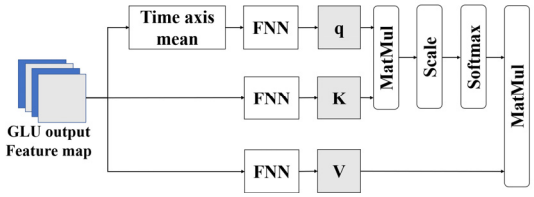


Fig. 3. (Color available online) Diagram of the proposed self-attention block.

log-Mel spectrogram으로 나타낸 그림이다. (b)는 (a)에서 나타낸 데이터가 GLU를 통해 나온 특징맵을 시각화하기 위해 입력데이터의 크기로 재 샘플링 한 그림이다. (c)는 입력데이터 (a)를 GLU에서 나온 특징맵인 (b)에 적용했을 때의 결과를 나타낸다. 위의 그림을 통해 입력데이터의 중요한 정보 값은 유지하고 배경잡음같이 불필요한 정보는 제거하는 것을 알 수 있다. 각 그림의 가로축은 시간, 세로축은 주파수를 나타낸다. 그림의 색상은 색이 밝을수록 값이 크고, 어두울수록 값이 작다.

본 논문에서는 GLU에서 나온 특징맵에 max-pooling 층을 적용하여 과적합 문제를 방지한다. Fig. 2는 위의 방법들을 사용한 GLU-block을 나타낸다. 여러 GLU-block을 통해 입력데이터의 중요 정보에만 가중치를 부여하여 배경잡음에 강인한 특징을 추출할 수 있다.

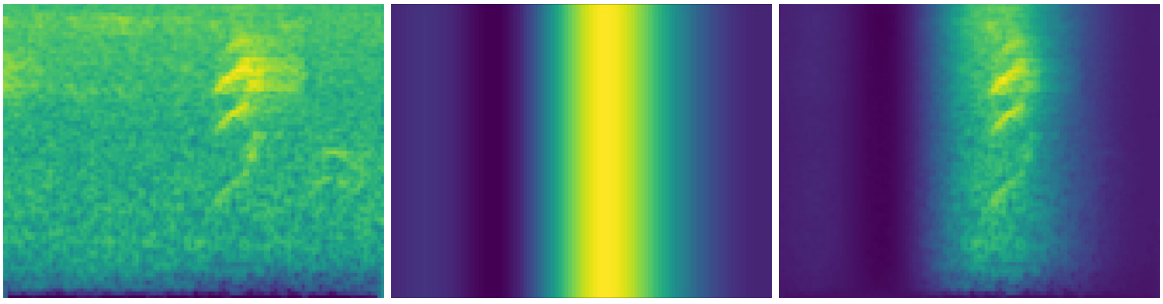
2.2 자가주의집중기법(Self-attention)

Attention 기법은 전체 입력 시퀀스로부터 정보를 통합하는 신경망 구조의 한 종류이다. 이때 정보를 통합하는 과정에서 입력데이터의 프레임별 중요도를 판별한다. 최근에 발표한 논문에서는 CNN과 RNN 같은 네트워크를 거치지 않고 입력데이터의 자체적으로 정보를 통합하는 과정에서 프레임별 중요도를 판별하는 방법을 제안했다.[10]

본 논문에서는 위의 방법에 착안하여 GLU-block을 통해 나온 정보를 사전정보 없이 프레임별 중요도를 판별해 인식률을 높일 수 있는 자가주의집중기법을 사용하였다.

자가주의집중기법은 q(Query), K(Key), V(Value) 총 3개의 변수가 필요하다. 그래서 Fig. 3처럼 입력데이터를 각각 d 차원의 q 벡터, K 행렬, V 행렬을 생성한다. 여기서 q는 중요도를 판별하고자 하는 벡터다. K 행렬은 q와 연산함으로써 프레임간 유사도를 계산하는 역할을 한다. V 행렬은 q와 K의 연산으로 나온 유사도의 softmax 확률값을 통해 주의집중 값을 도출하는 역할을 한다. 제안하는 방법에서는 시간축으로 자가주의집중기법을 적용해 프레임별 중요도 판별을 제안하기 때문에 q는 시간 축으로 평균을 취한 뒤 벡터를 생성한다. 본 논문에서 d는 128로 설정하였다.

$$\alpha = \mathrm{softmax}\left(\frac{q \times K^{T}}{\sqrt{d}}\right) \times V . \quad (2)$$



(a) Input data before apply self-attention (b) Self-attention of input data (c) Output data after self-attention layer

Fig. 4. (Color available online) Diagram when Self-attention is applied to data.

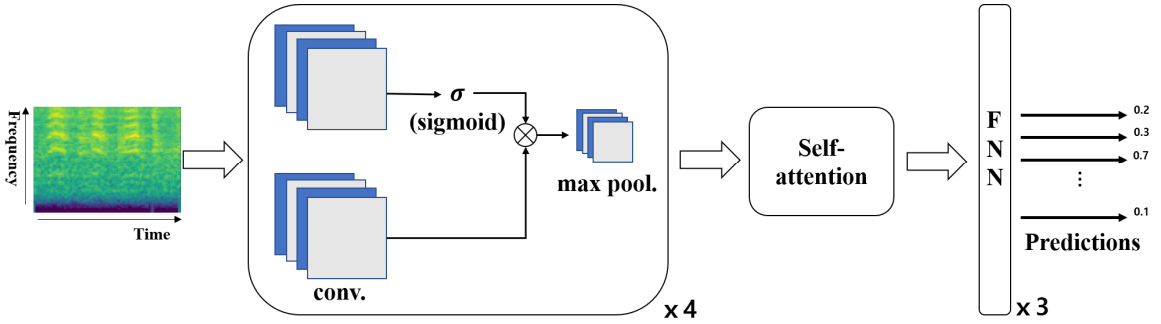


Fig. 5. (Color available online) Diagram of the proposed method.

Eq. (2)에서 α는 자가주의집중 값을 의미한다. q는 어떤 정보와 관련되어 있는지 찾기 위해 Eq. (2)처럼 K와 내적 연산을 한다. 그 후 softmax를 적용해, q가 K와의 연관성을 계산한 뒤 그 값들을 확률값으로 만들어 준다. Eq. (2)에서 d는 q, K, V의 차원을 의미하고 q와 K의 값이 커지면 제대로 된 자가주의집중 값을 도출할 수 없으므로 $\sqrt{d}$로 나눈다. 위 식을 가지고 q가 K의 값 중에서 확률을 통해 데이터 간 연관성을 가지는지 알 수 있는 주의집중 값을 구할 수 있다. q와 K로 나온 확률값을 V에 곱해서 주의집중 값을 계산한다. 최종적으로 자가주의집중기법을 통해 d의 값으로 설정한 128차원의 벡터가 나온다.

Fig. 4(a)는 실제 동물 울음소리의 입력데이터를 log-Mel spectrogram으로 나타낸 그림이다. (b)는 (a)에서 나타낸 소리 데이터가 GLU를 통해 나온 특징맵의 자가주의집중 값을 시각화하기 위해 입력데이터의 크기로 재샘플링한 그림이다. 각 그림의 가로축은 시간, 세로축은 주파수를 나타낸다. 그림의 색상은 색이 밝을수록 값이 크고, 어두울수록 값이 작다. (c)는 입력데이터 (a)를 입력받은 GLU 모듈의 최종 출력 특징맵을 자가주의집중 Eq. (2)의 softmax 결괏값을 (b) 적용했을 때의 결과를 나타낸다. 위의 그림을 통해 입력데이터에서 동물 울음소리같이 중요한 정보는 최대한 유지하고 그 외에 배경잡음같이 필요없는 정보의 값은 무시한다는 것을 알 수 있다. 최종적으로 (b)가 다음 층에 입력이 된다.

2.3 모델 구조

본 논문에서 제안하는 알고리즘 구조는 Fig. 5처럼 4개의 GLU-block, 시간 축으로의 자가주의집중 그리고 3개의 FNN을 사용한다. 먼저 4개의 GLU-block을 통해 중요한 정보의 값은 살리고 불필요한 정보를 최대한 줄인다. GLU 사용 시 kernel stride는 1로 설정했다. 이때 과적합 문제를 방지하기 위해 GLU를 거친 뒤 batch-normalization[11]을 적용했다. 그리고 시간 축을 중심으로 자가주의집중을 적용해 데이터의 프레임별 중요도를 판별해 원하는 정보를 더욱 수월하게 찾을 수 있게 모델을 설계했다. 실험에서 쓰인 q, K, V의 차원은 128로 설정했다.

III. 실 험

3.1 데이터베이스

본 실험에 사용하는 데이터는 국내에 서식하고 있는 양서류 8종과 조류 6종의 울음소리를 자체 수집을 통해 구성했다(44.1 kHz, 단일 채널, 16 bit-resolution). 양서류 종은 금개구리(pelophylax chosenicus), 맹꽁이(kaloula borealis), 무당개구리(bombina orientalis), 북방산개구리(rana dybowskii), 수원청개구리(hyla suweonensis), 옴개구리(glandirana rugosa), 참개구리(pelophylax nigromaculatus), 청개구리(hyla japonica)로 총 8종으로 구성했다. 조류는 긴꼬리딱새(terpsiphone atrocaudata), 동박새(zosterops japonicus), 매(falco peregrinus), 팔색조(pitta nympha), 호랑지빠귀(zoothera dauma), 흑비둘기(columba janthina)로 총 6종으로 구성했다. Table 1는 14종의 개체목록이 녹음된 신호의 데이터 개수를 보여준다. Train, Valid, Test의 데이터 총 길이는 각각 2.90 h, 1.82 h, 1.17 h이다. 모든 데이터의 길이는 1 s로 구성했다.

3.2 실험방법

실험을 위해 1차원 데이터인 오디오 샘플을 1 s씩 자른 뒤 샘플링 속도는 22050 Hz, 윈도우 사이즈는 36.3 ms, hop size는 9 ms, FFT 크기는 2048, Mel-filter는 80개로 설정해 크기가 111 × 80인 log-Mel spectrogram으로 변환했다. 자가주의집중 기법을 적용할 때 d는 128로 설정했다. 네트워크학습에서 기울기 소실문제를 줄이기 위해 활성화 함수로는 ReLU 함수를 사용했고, epoch은 200, batch size는 64로 설정했다. 모델 훈련 시 검증데이터 음원에 대해 평균 정확도가 제일 높을 때의 모델의 파라미터 값을 가지고 실험을 진행했다. 또한, 종들의 데이터 개수가 불균형하기 때문에 모델의 종별 정확도에 대한 척도를 측정하기 위해 F1-score도 함께 확인했다.

먼저 본 논문의 베이스라인인 4개의 CNN과 3개의 FNN으로 구성한 모델과 제안하는 구조인 4개의 GLU-block과 시간 축에서의 자가주의집중 모듈, 3개의 FNN으로 구성한 모델을 실험했다. 베이스라인은 AlexNet을 기반으로 설계하였다. 또한, CNN과 GLU의 성능 차이를 비교하기 위해 베이스라인에서 사용한 모든 CNN 층을 GLU-block으로 바꿔서 실험했다. 마지막으로 객관적인 성능 비교를 위해 기존 분류에 쓰이는 모델인 5개의 CNN과 3개의 FNN을 사용한 AlexNet 구조도 함께 실험했다.[12,13]

추가로 잡음이 많이 들어간 데이터에서도 중요 정보를 추출할 수 있는지 확인하기 위해 자연환경에서 발생할 수 있는 잡음인 바람, 비, 폭우 그리고 강 소리를 섞은 데이터도 실험하였다. 잡음 데이터의 신호

Table 1. List of amphibian and bird classes and the number of train set, validation set and test set.

Species	Train	Valid	Test	Total
pelophylax chosenicus	372	102	104	578
kaloula borealis	1199	479	499	2177
bombina orientalis	2104	1277	567	3948
rana dybowskii	754	530	231	1515
hyla suweonensis	1348	1447	696	3491
glandirana rugosa	493	415	161	1069
pelophylax nigromaculatus	934	959	770	2663
hyla japonica	1264	304	254	1822
terpsiphone atrocaudata	174	151	155	480
zosterops japonicus	290	79	64	433
falco peregrinus	216	109	90	415
pitta nympha	559	406	250	1215
zoothera dauma	599	249	315	1163
columba janthina	117	61	57	235
Total	10423	6568	4213	21204

Table 2. Detail parameters of proposed model.

Layer {kernel size, max pooling size}		# kernel or # dimension	Output size (# T × F × C)
Input of Log-mel spectrogram		-	111 × 80 × 1
1st GLU-block {3x3, 2x2}		16	55 × 40 × 16
2nd GLU-block {3x3, 2x2}		32	27 × 20 × 32
3rd GLU-block {3x3, 2x2}		64	13 × 10 × 64
4th GLU-block {3x3, 2x2}		128	6 × 5 × 128
Reshape		-	6 × 640
Self-attention	**q** : 1 × 128	128	128
	V : 6 × 128		
	K : 6 × 128		
FNN, ReLU		64	64
FNN, ReLU		32	32
FNN		# classes	# classes

대 잡음 비(Signal to Noise Ratio, SNR)는 0 dB로 설정해 성능을 측정하였다.

Table 2은 제안한 모델의 세부 파라미터를 보여준다. 여기서 T, F, C는 각각 time, frequency, channel을 의미한다. 실험은 Python과 Tensorflow2.0 패키지를 사용했고, 훈련시간은 모델당 약 1.5시간이 소요되었다.

IV. 결과 및 토의

Tables 3과 4는 각각 조류, 양서류를 입력데이터로 넣었을 때 베이스라인과 AlexNet, 베이스라인에서 CNN대신 GLU-block을 넣은 모델 그리고 마지막으로 본 논문에서 제안 방법인 GLU와 자가주의집중기법을 사용한 모델을 사용해 진행한 실험 결과를 정리한 것이다. 여기서 GLU와 FNN 앞에 있는 숫자는 각각 GLU-block과 FNN의 층의 수를 의미한다.

실험 결과 Tables 3과 4에서도 볼 수 있듯이 본 논문에서 제안한 구조가 조류를 분류할 때 90.72 %, 양서류를 분류할 때 93 %로 베이스라인이 조류를 분류할 때 87 %, 양서류를 분류할 때 90 %로 제안하는 구조가 약 3 % 정도 더 높은 정확도를 보였다. 그리고 베이스라인에서 CNN대신 GLU를 활용한 구조가 조류를 분류할 때 89 %로 베이스라인의 분류율인 87 %보다 2 % 높지만, 양서류를 분류할 때 88 %로 베이스라인의 분류율인 90 %로 더 떨어졌다. 하지만 양서류를 실험할 때 F1-score를 비교하면 베이스라인에서 CNN대신 GLU-block을 활용한 구조의 F1-score가 81 %로 베이스라인의 F1-score인 80 %보다 더 높은 것으로 나타났다. 이를 통해 CNN대신 GLU를 사용했을 때 입력데이터의 중요 정보에 가중치를 부여해 배경 잡음에 강인한 특징을 추출하기 때문에 F1- score 기준으로 더 나은 성능을 보인다는 결론에 도달할 수 있다. 또 기존에 제안된 구조인 AlexNet 성능을 비교했을 때 조류의 평균 정확도와 F1-score는 각각 84 %, 73 %이고, 양서류의 평균 정확도와 F1-score는 각각 87 %, 75 %로 제안한 알고리즘에서 조류의 평균 정확도와 F1-score, 양서류의 평균 정확도와 F1-score와 비교했을 때 성능이 현저히 떨어지는 것으로 나타났다.

Tables 5와 6은 각각 조류와 양서류 데이터에 잡음을 섞었을 때 베이스라인과 AlexNet, 베이스라인에서 CNN대신 GLU-block을 넣은 모델 그리고 마지막으로 본 논문에서 제안 방법인 GLU와 자가주의집중기법을 사용한 모델을 사용해 진행한 실험 결과를 정리한 것이다. 실험 결과 본 논문에서 제안한 구조가 잡음이 섞인 조류와 양서류 데이터를 분류할 때 각각 85 %, 92 %로 베이스라인으로 잡음이 섞인 조류와 양서류를 분류할 때 각각 80 %, 85 %로 제안하는 구조가 약 5 % ~ 7 % 정도 더 높은 정확도를 보였다. 또 기존에 제안된 구조인 AlexNet 성능을 비교했을 때 조류의 평균 정확도와 F1-score는 각각 78 %, 67 %

Table 3. Experiment results of bird sound according to each architecture.

Architecture	terpsiphone atrocaudata	zosterops japonicus	falco peregrinus	pitta nympha	zoothera dauma	columba janthina	Avg. Acc	F1-score
AlexNet	74 %	50 %	30 %	93 %	100 %	98 %	84 %	73 %
Baseline	87 %	89 %	32 %	92 %	100 %	83 %	87 %	82 %
4GLU-3FNN	94 %	88 %	32 %	96 %	100 %	84 %	89 %	84 %
Proposed Method	99 %	92 %	60 %	97 %	100 %	56 %	91 %	87 %

Table 4. Experiment results of amphibian sound according to each architecture.

Architecture	pelophylax chosenicus	kaloula borealis	bombina orientalis	rana dybowskii	hyla suweonensis	glandirana rugosa	pelophylax nigro-maculatus	hyla japonica	Avg. Acc	F1-score
AlexNet	38 %	100 %	69 %	38 %	100 %	71 %	98 %	100 %	87 %	75 %
Baseline	49 %	100 %	88 %	49 %	99 %	61 %	99 %	100 %	90 %	80 %
4GLU-3FNN	71 %	100 %	67 %	46 %	93 %	100 %	100 %	100 %	88 %	81 %
Proposed Method	66 %	99 %	84 %	51 %	100 %	97 %	99 %	100 %	93 %	85 %

Table 5. Experiment results of bird sound with noise signal according to each architecture.

Architecture	terpsiphone atrocaudata	zosterops japonicus	falco peregrinus	pitta nympha	zoothera dauma	columba janthina	Avg. Acc	F1-score
AlexNet	85 %	73 %	49 %	84 %	98 %	19 %	78 %	67 %
Baseline	97 %	75 %	40 %	86 %	99 %	14 %	80 %	70 %
4GLU-3FNN	98 %	86 %	53 %	89 %	100 %	14 %	83 %	74 %
Proposed Method	95 %	89 %	69 %	91 %	100 %	14 %	85 %	76 %

Table 6. Experiment results of amphibian sound with noise signal according to each architecture.

Architecture	pelophylax chosenicus	kaloula borealis	bombina orientalis	rana dybowskii	hyla suweon-ensis	glandirana rugosa	pelophylax nigro-maculatus	hyla japonica	Avg. Acc	F1-score
AlexNet	33 %	95 %	87 %	44 %	100 %	45 %	96 %	100 %	87 %	75 %
Baseline	36 %	91 %	64 %	50 %	100 %	91 %	95 %	100 %	85 %	76 %
4GLU-3FNN	94 %	98 %	91 %	64 %	100 %	28 %	99 %	100 %	92 %	82 %
Proposed Method	84 %	99 %	89 %	56 %	100 %	48 %	98 %	100 %	92 %	83 %

이고, 양서류의 평균 정확도와 F1-score는 각각 87 %, 75 %로 제안한 알고리즘에서 조류의 평균 정확도와 F1-score, 양서류의 평균 정확도와 F1-score와 비교했을 때 성능이 제안한 구조보다 좋지 않았다. 이를 통해 제안한 구조가 잡음에 강인한 특징을 추출할 수 있다는 결론에 도달할 수 있다.

V. 결 론

본 논문에서는 GLU와 자가주의집중기법을 활용해 동물 울음소리데이터에 배경잡음같이 불필요한 정보를 최대한 제거해 구별 정확도를 높였다. 또한, 보편적으로 쓰이고 있는 알고리즘과 논문에서 제안한 구조의 차이점을 비교 및 분석했다. 그 결과 베이스라인보다 본 논문에서 제안한 구조를 사용한 구조의 분류 정확도가 약 2 %~3 %가량 더 높은 것을 확인했다. 또한, 기존에 쓰이는 알고리즘과 비교했을 때 분류 정확도는 약 6 % 차이가 나는 것을 보였고 잡음에도 강인한 특징을 추출할 수 있음을 확인했다. 추후 연구에서는 동물 종의 범위를 넓히고, 수집한 데이터를 통해 어느 시점에서 어느 종의 울음소리가 있는지 분리 및 분류하는 연구를 진행할 예정이다.

감사의 글

본 결과물은 환경부의 재원으로 한국환경산업기술원의 환경정책기반공공기술개발사업의 지원을 받아 연구되었습니다(2017000210001).

References

1. K. Ko, J. Park, D. K. Han, and H. Ko, “Channel and frequency attention module for diverse animal sound classification,” Proc. IEICE Trans. on Information and Systems, **E102.D**, 2615-2618 (2019).
2. J. Strout, B. Rogan, S. M. M. Seyednezhad, K. Smart, M. Bush, and E. Ribeiro, “Anuran call classification with deep learning,” Proc. IEEE ICASSP. 2662-2665 (2017).
3. X. Dong, N. Yan, and Y. Wei, “Insect sound recognition based on convolutional neural network,” Proc, IEEE 3rd ICIVC. 855-859 (2018).
4. J. Salamon, J. P. Bello, A. Farnsworth, and S. Kelling, “Fusing shallow and deep learning for bioacoustic bird species classification,” Proc. IEEE ICASSP. 141-145 (2017).
5. K. Ko, S. Park, and H. Ko, “Convolutional feature vectors and support vector machine for animal sound classification,” IEEE 40th EMBC. 376-379 (2018).
6. Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, “Large-scale weakly supervised audio classification using gated convolutional neural network,” Proc. IEEE ICASSP. 121-125 (2018).

7. J. Yan, Y. Song, W. Guo, L.-R. Dai, I. McLoughlin, and L. Chen, “A region based attention method for weakly supervised sound event detection and classification,” Proc. IEEE ICASSP. 755-759 (2019).
8. K. Miyazaki, T. Komatsu, T. Hayashi, S. Watanabe, T. Toda, and K. Takeda, “Weakly-supervised sound event detection with self-attention,” Proc. IEEE ICASSP. 66-70 (2020).
9. Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” Proc. ICML. 933-941 (2017).
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in NPS30. 5998- 6008 (2017).
11. S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” arXiv preprint arXiv:1502.03167 (2015).
12. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in NPS25. 1097-1105 (2012).
13. K. Ko, S. Park, and H. Ko, “Convolutional neural network based amphibian sound classification using covariance and modulogram” (in Korean), J. Acoust. Soc. Kr. **37**, 60-65 (2018).

저자 약력

▸ 김 정 민 (Jungmin Kim)

2019년 2월 : 제주대학교 해양시스템공학과 학사 취득
2019년 3월 ~ 현재 : 고려대학교 전기전자공학과 석사과정

▸ 이 영 로 (Younglo Lee)

2015년 2월 : 고려대학교 전기전자전파공학부 학사 취득
2015년 3월 ~ 현재 : 고려대학교 전기전자공학과 석박사통합과정

▸ 김 동 현 (Donghyeon Kim)

2018년 2월 : 영남대학교 전기공학과 학사 취득
2018년 3월 ~ 현재 : 고려대학교 전기전자공학과 석박사통합과정

▸ 고 한 석 (Hanseok Ko)

1982년 : Carnegie-Mellon Univ. 전기공학 공학사 취득
1988년 : Johns Hopkins Univ. 전자공학 공학석사 취득
1992년 : Catholic Univ. of America 전자공학 공학박사 취득
1994년 ~ 현재 : 고려대학교 전기전자공학과 교수