SED를 활용한 사고 대응 시스템

2020 소프트웨어응용

2018580001 강창구
2018920059 허정우

목차

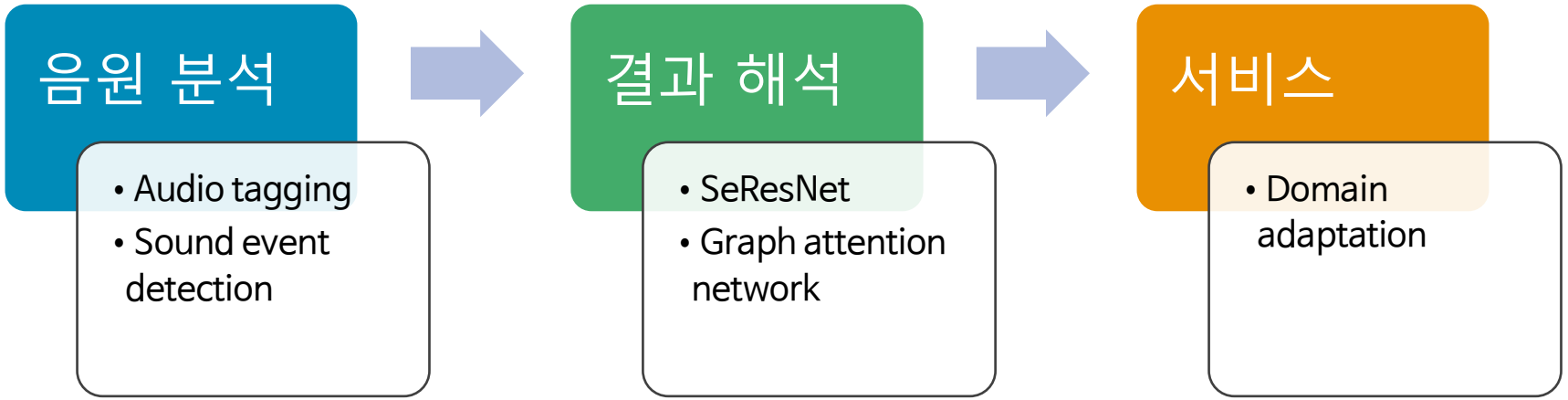
소리를 활용한 신속한 사고 대응 시스템

SED* 를 활용한 사고 대응 시스템

- SED를 활용한 사고 대응 시스템은 주변의 소리 정보를 분석하고, 분석된 정보를 해석하여 해당 지역의 위험도를 산출하는 시스템입니다.
- 본 시스템을 활용하면 기존 CCTV에 비해 적은 비용을 지불하고도 효율적인 도시 안전망을 구축할 수 있습니다.

* SED : Sound Event Detection 의 약자

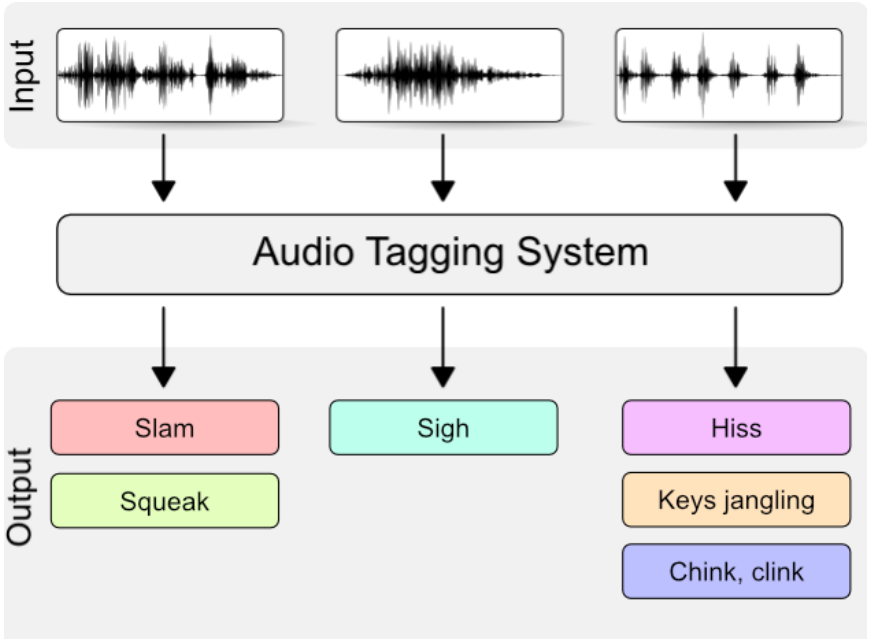
프로젝트 로드맵


음원 분석
• Audio tagging
• Sound event detection
결과 해석
• SeResNet
• Graph attention network
서비스
• Domain adaptation

Audio Tagging / SED

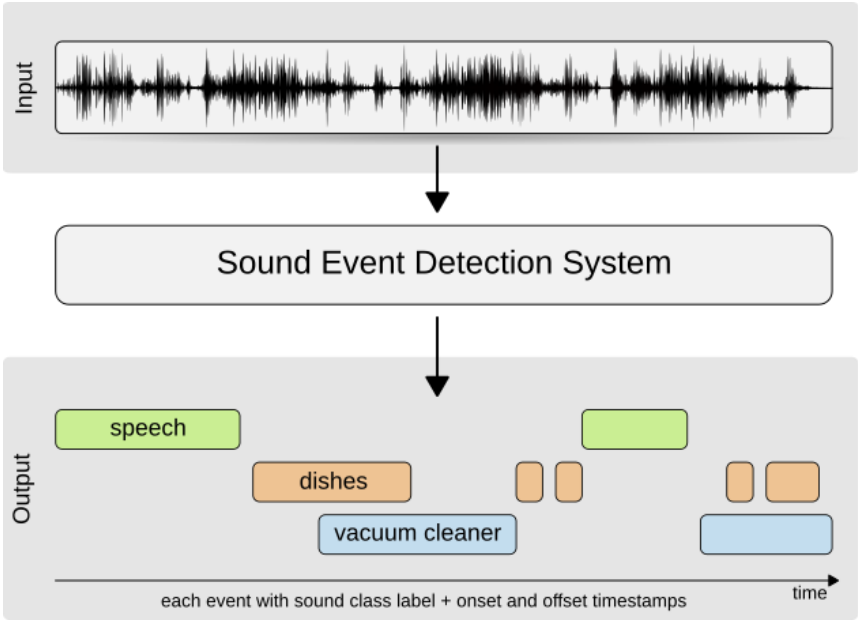
Audio Tagging

주어진 입력이 무엇을 녹음했는지 식별하는 과제



SED

무엇을 녹음했는지 식별한 후, 어느 부분에 녹음이 되었는지 식별하는 과제

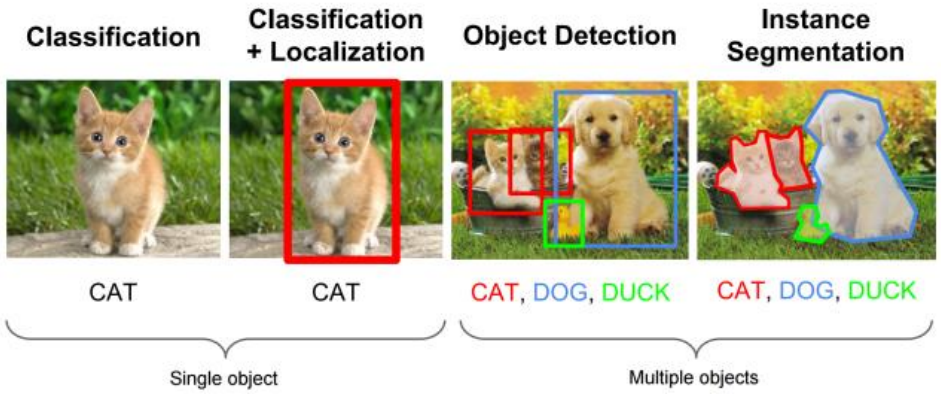


* 위 이미지는 2019 DCASE TASK2, 2020 DCASE TASK5 DESCRIPTION에서 가져옴

프로젝트 관련 연구

Sound Event Detection (SED) 구현 - 1.CNN

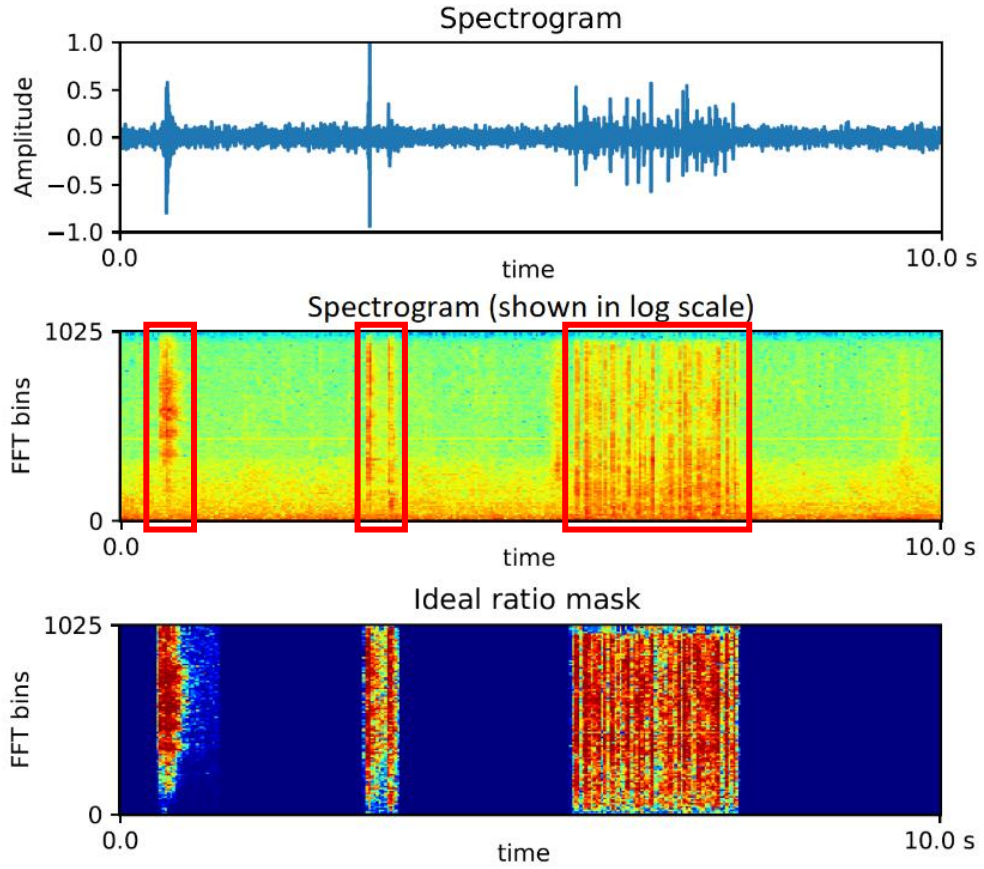
Image



audio의 1차원 waveform 에 STFT 를 적용하여
시간-주파수 2차원 mel-spectrogram 으로 변환 후,

배경음과 다른 특징을 가진 소리를 검출하기 위해
image 의 localization, segmentation 을 적용

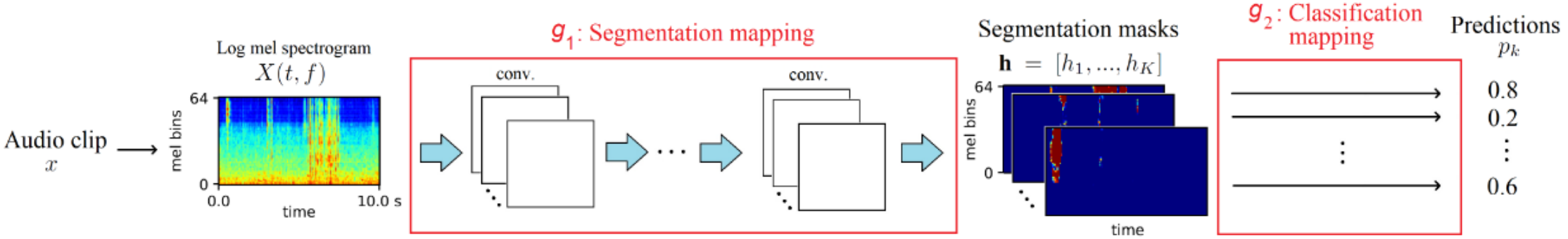
Sound

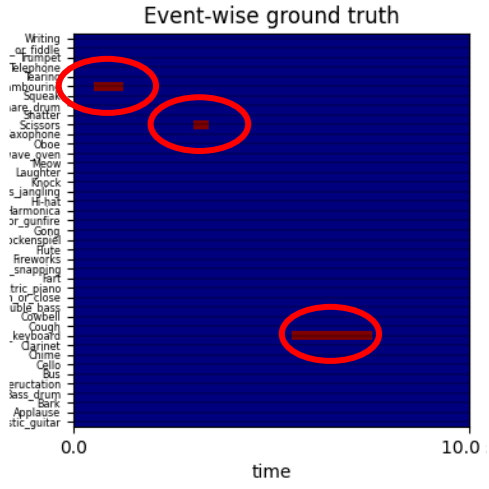


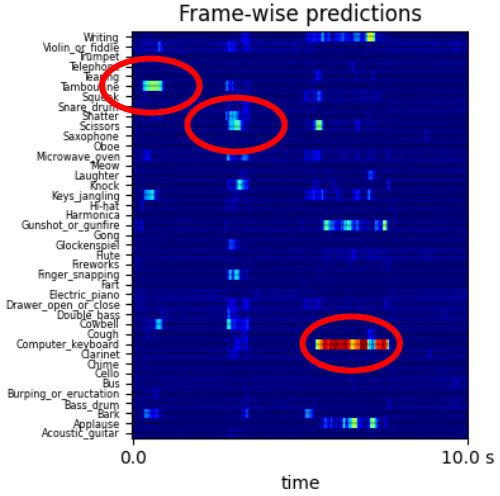
프로젝트 관련 연구

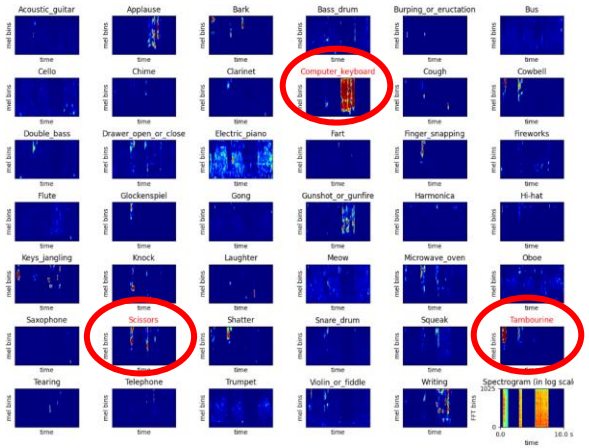
Sound Event Detection (SED) 구현 - 1.CNN

시간 축에 대한 정보를 통합하면서 손실 발생, RNN층 결합 시 성능 향상 기대 <- GAP/GMP/GWRP









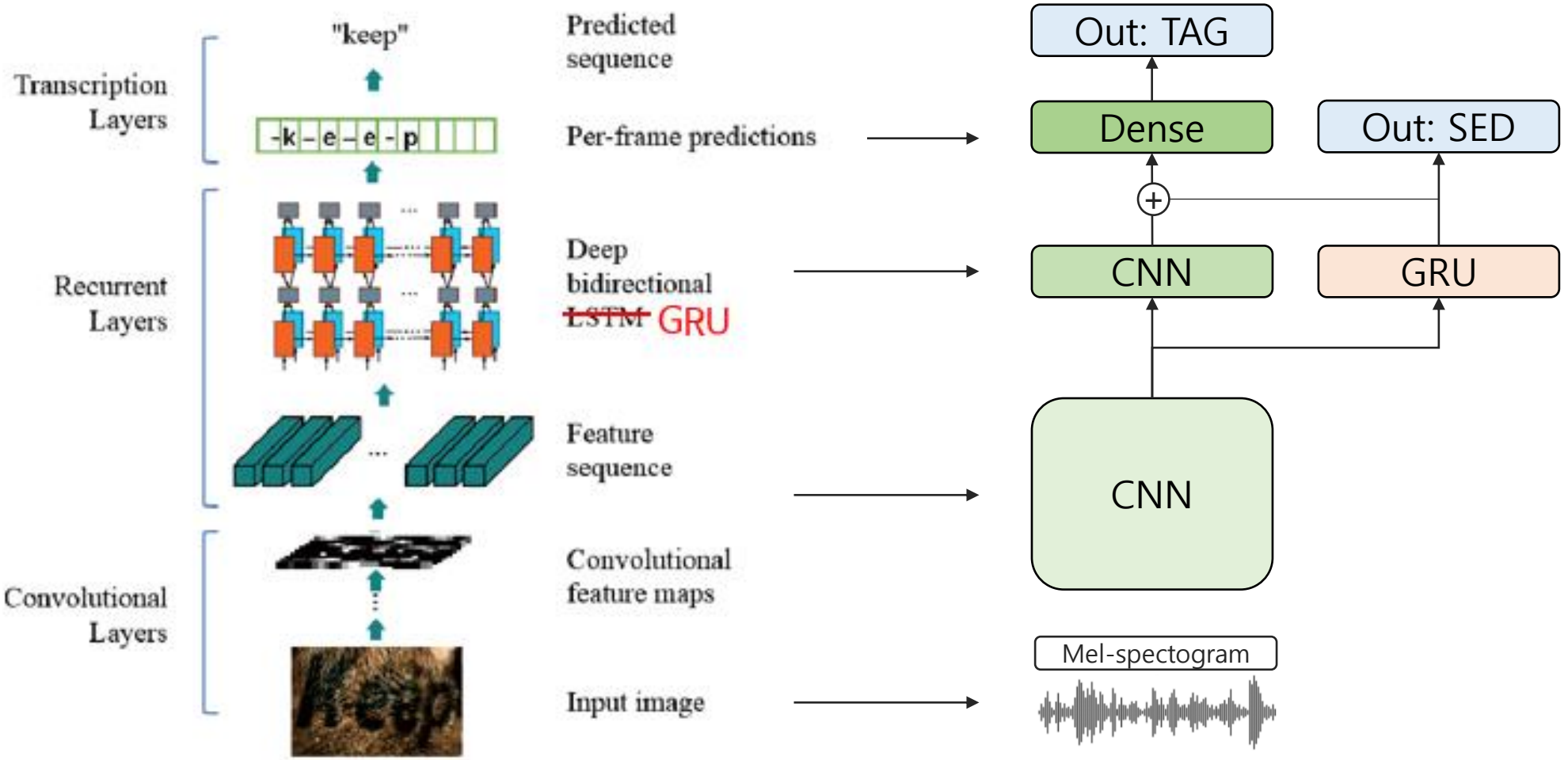
F1-SCORE, AUC AND MAP OF AUDIO TAGGING AT DIFFERENT SNRS.

Algorithms	20 dB			10 dB			0 dB		
	F1	AUC	mAP	F1	AUC	mAP	F1	AUC	mAP
DNN [55]	0.439	0.885	0.468	0.396	0.861	0.402	0.331	0.810	0.314
WLD CNN [37]	0.498	0.777	0.498	0.524	0.794	0.526	0.528	0.815	0.535
FrameCNN [34]	0.581	0.899	0.587	0.543	0.883	0.526	0.484	0.850	0.439
Attention [39]	**0.714**	0.922	**0.755**	**0.690**	0.907	**0.729**	**0.612**	0.875	**0.643**
GMP	0.435	0.818	0.475	0.406	0.801	0.440	0.373	0.773	0.389
GAP	0.529	0.934	0.623	0.467	0.914	0.555	0.385	0.877	0.442
GWRP	0.635	**0.955**	0.753	0.604	**0.942**	0.696	0.534	**0.915**	0.596

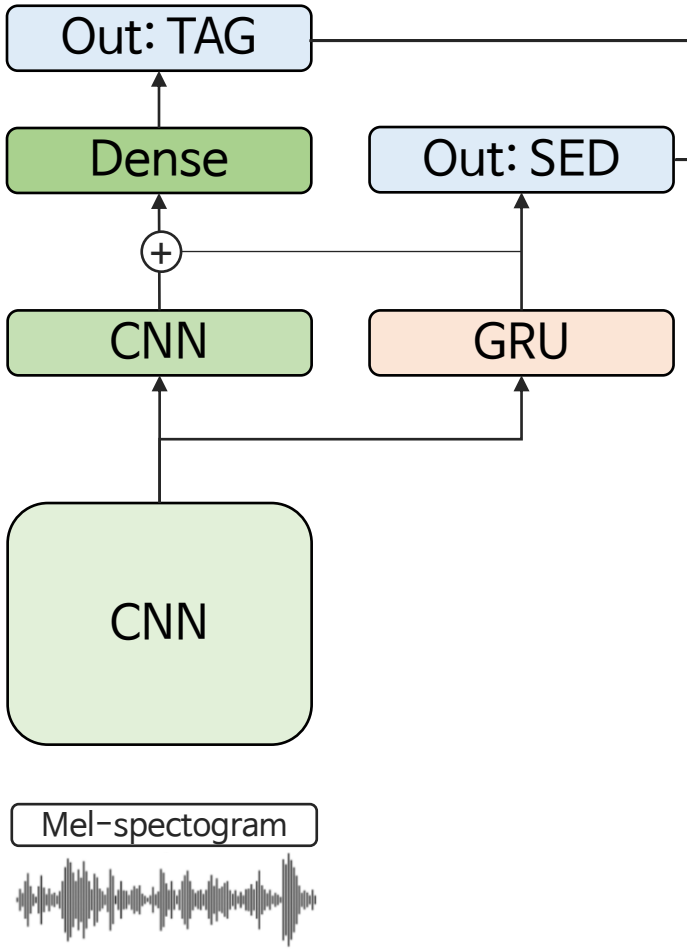
시각적인 확인/평가 척도 계산 시, 성능 낮음 확인 가능

프로젝트 관련 연구

Sound Event Detection (SED) 구현 - 2.CRNN



Sound Event Detection (SED) 구현 - 2.CRNN



Architecture	TAG *lwlrap*	SED	
		F1	ER
Baseline1	69.41	79.62	0.2926
Baseline2	70.62	78.61	30.85

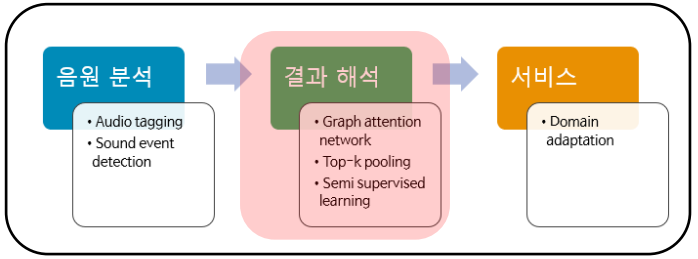
$lwlrap^*$ = label-weighted label-ranking average precision

$$\frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} \frac{1}{||y_i||_0} \sum_{j:y_{ij}=1} \frac{|\mathcal{L}_{ij}|}{\text{rank}_{ij}}$$

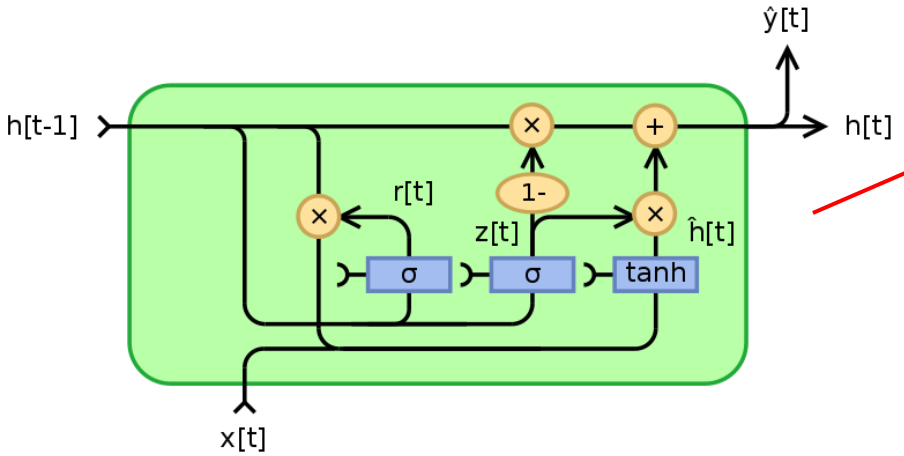
*lwlrp - https://scikit-learn.org/stable/modules/generated/sklearn.metrics.label_ranking_average_precision_score.html

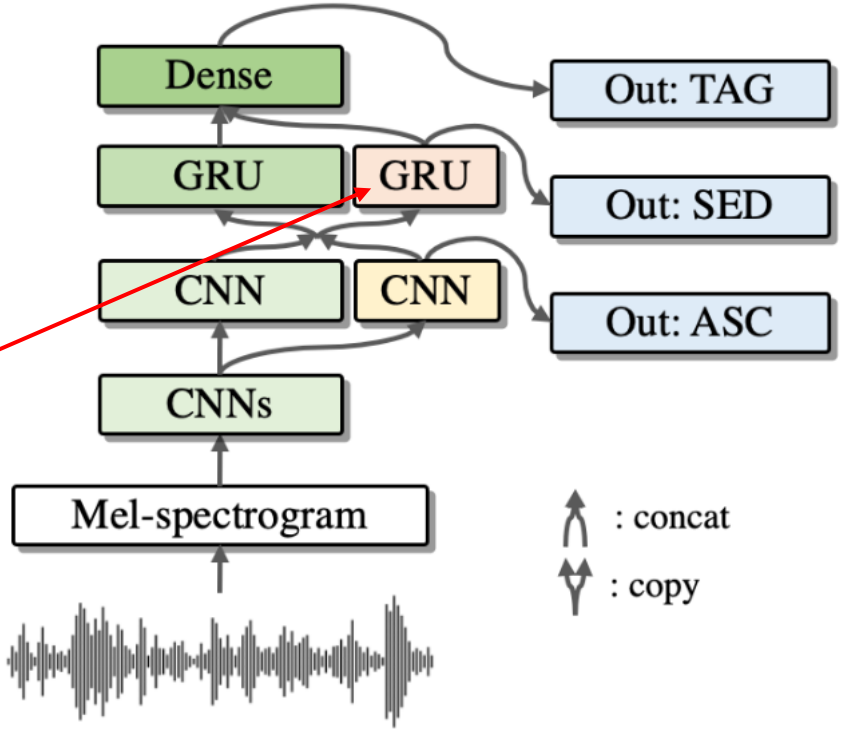
모델 관련 연구

DCASENet*



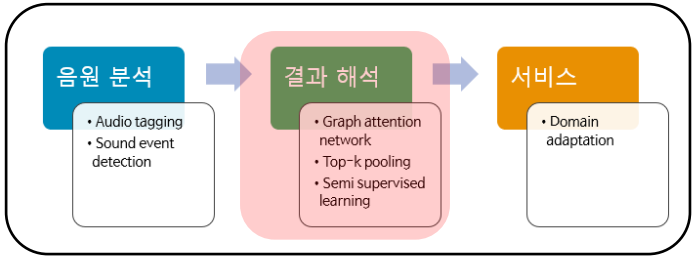
- DCASE의 3가지 과제(ASC, AT, SED)가 연관성 있음을 가정, 동시 수행하는 프레임워크
- 본 프로젝트도 세 과제와 연관성 있음을 가정, 음원 분석을 위해 사용



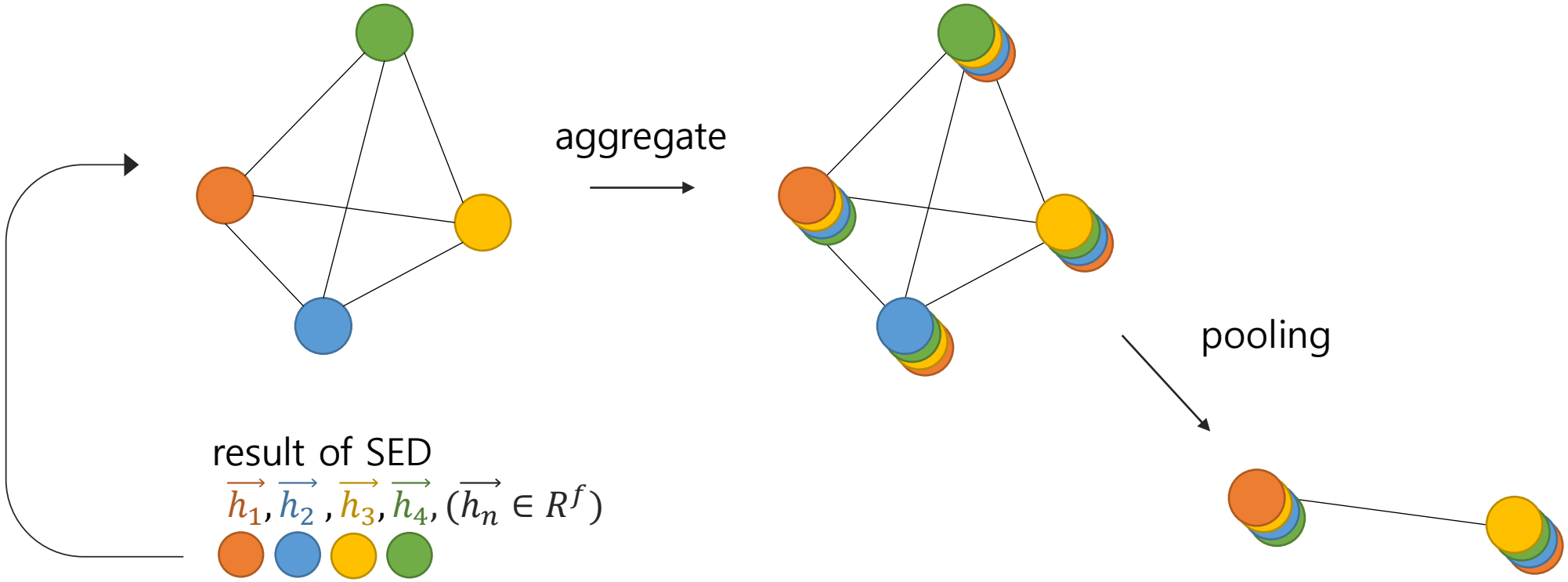


* Jung, J. W., Shim, H. J., Kim, J. H., & Yu, H. J. (2020). DCASENET: A joint pre-trained deep neural network for detecting and classifying acoustic scenes and events. *arXiv preprint arXiv:2009.09642*

Graph attention network(GAT*)



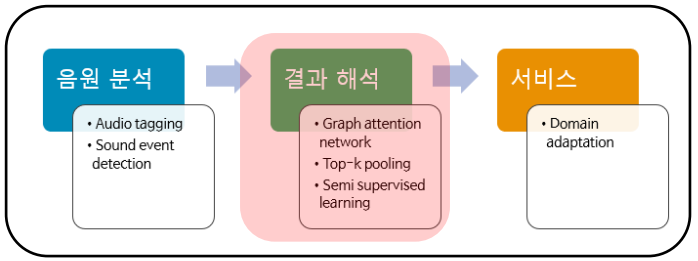
- GNN의 aggreagete 과정에서 타겟 노드의 feature를 Key, 인접 노드의 feature를 Query, Value로 설정하고 Self-Attention 으로 노드 사이의 관계 학습



* Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 'Graph attention networks.' In International Conference on Learning Representations (ICLR), 2018.

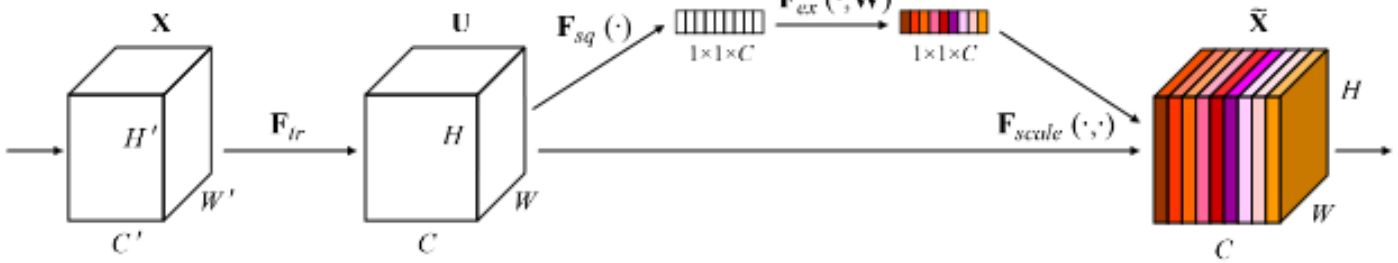
모델 관련 연구

Squeeze-and-excitation networks(SENet*)

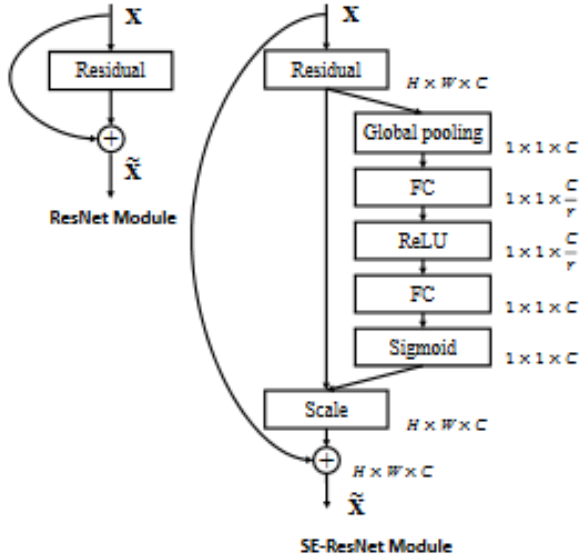


- 채널 간의 상호작용을 학습한 뒤, 해당 정보를 사용해 채널 단위로 새로운 가중치를 부여해 성능 향상

Squeeze-and-excitation Block



SE-ResNet Module



* Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

데이터셋

1. DCASE* Dataset

Scene classification

Audio tagging

SED

SED

2. Free Sound / YouTube

위험 이벤트 음원 파일

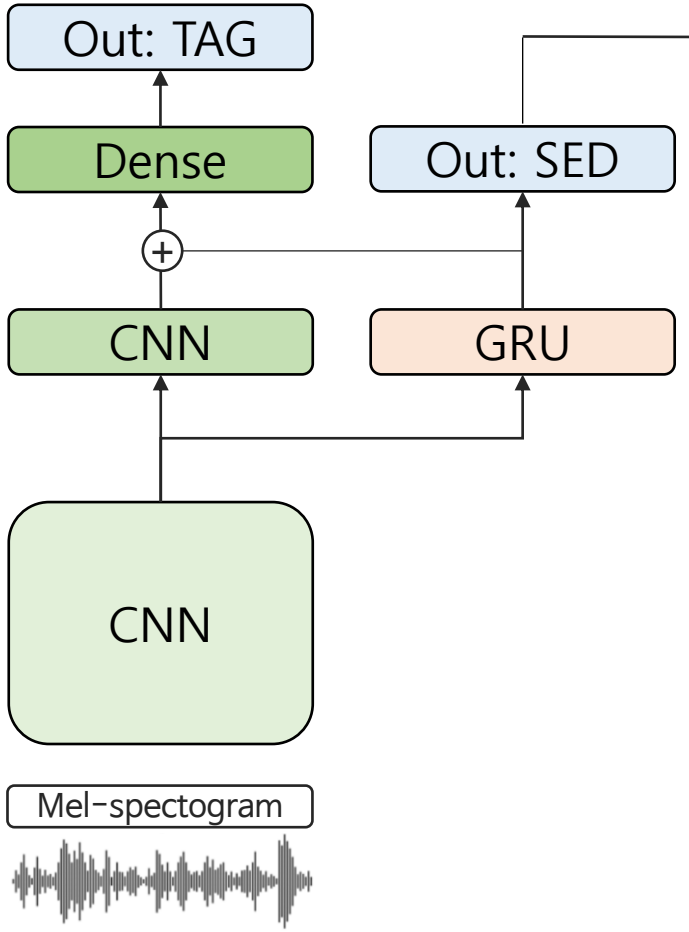
국내외 재난 영상 음원 파일

* DCASE: IEEE에서 주관하는 음향 신호 탐지 및 식별에 관한 챌린지

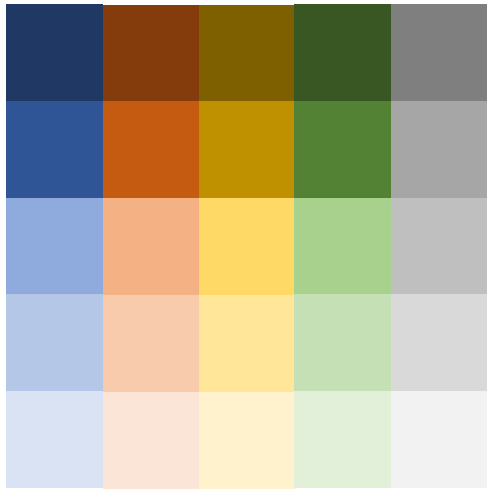
실험 구성 - 1. 음성 특징 추출

Mel Spectrogram Extractor

- sample rate : 44100kHz
- window length : 40ms
- hop length : 20ms
- fft size : 2048
- mel bins : 128



(batch_size, frame, feature)



feature

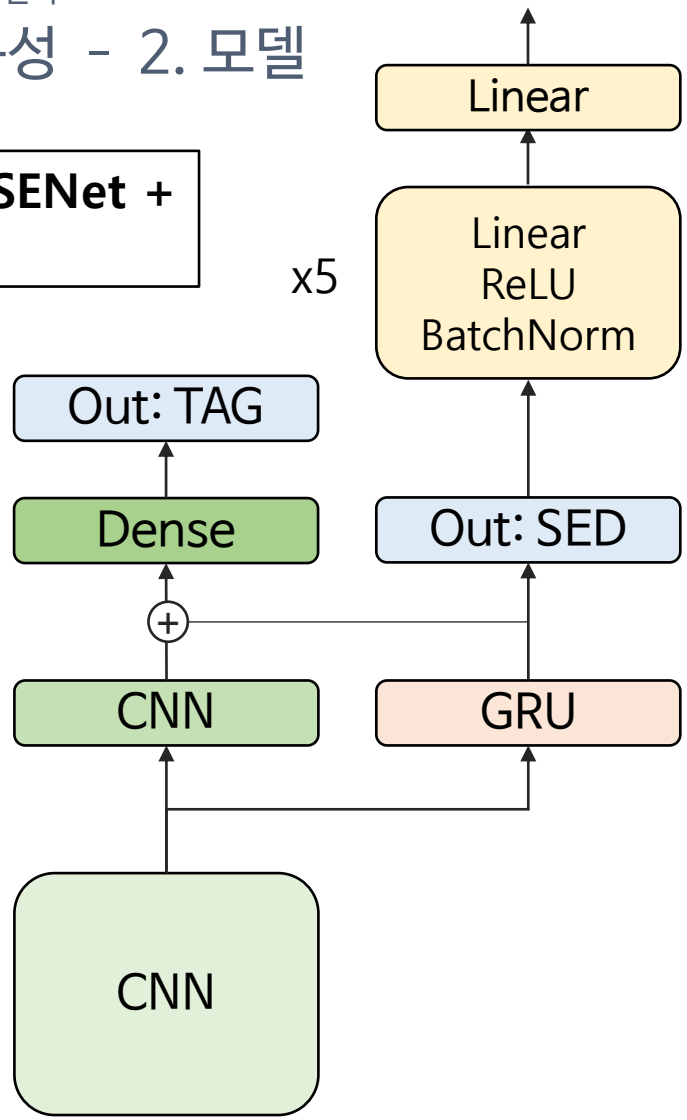
frame

pre-trained DcaseNet v3

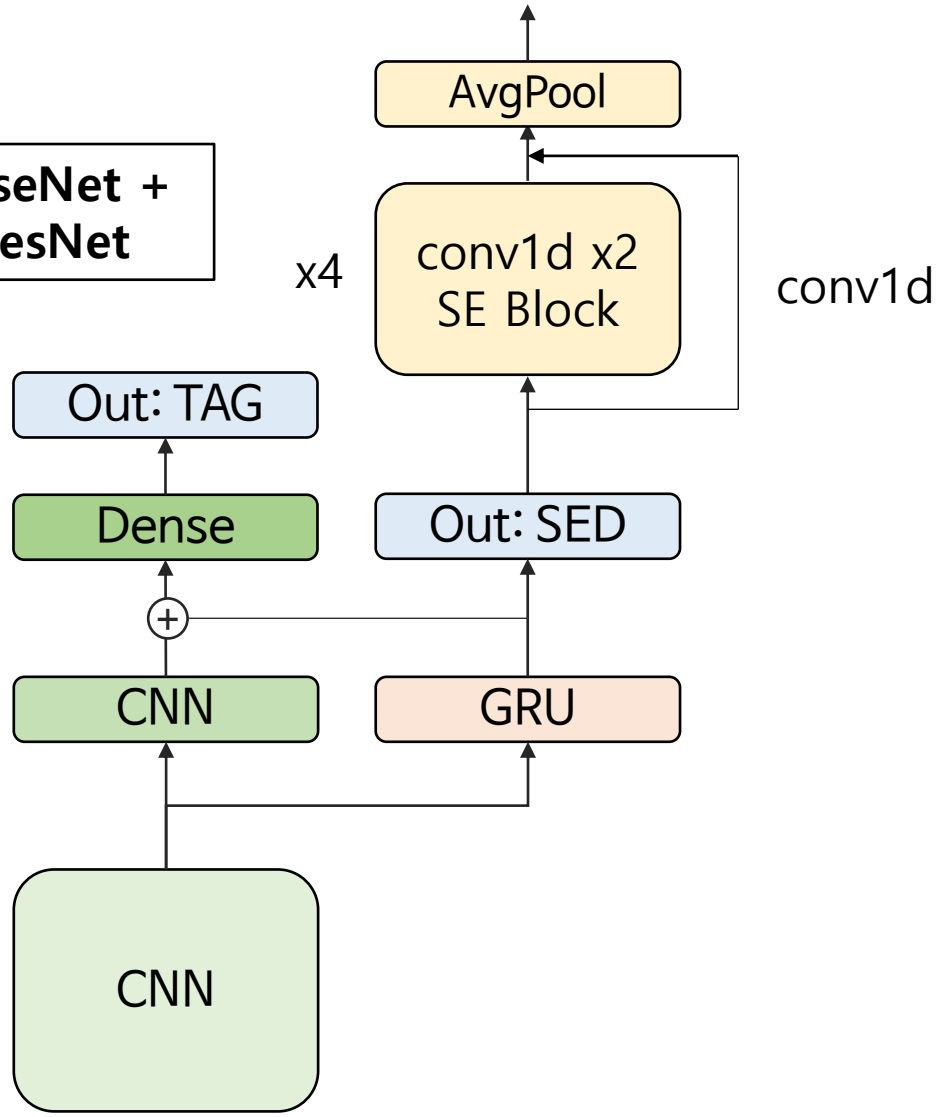
프레임별 음성특징 추출기로 사용

실험 구성 - 2. 모델

1. DCASENet + MLP

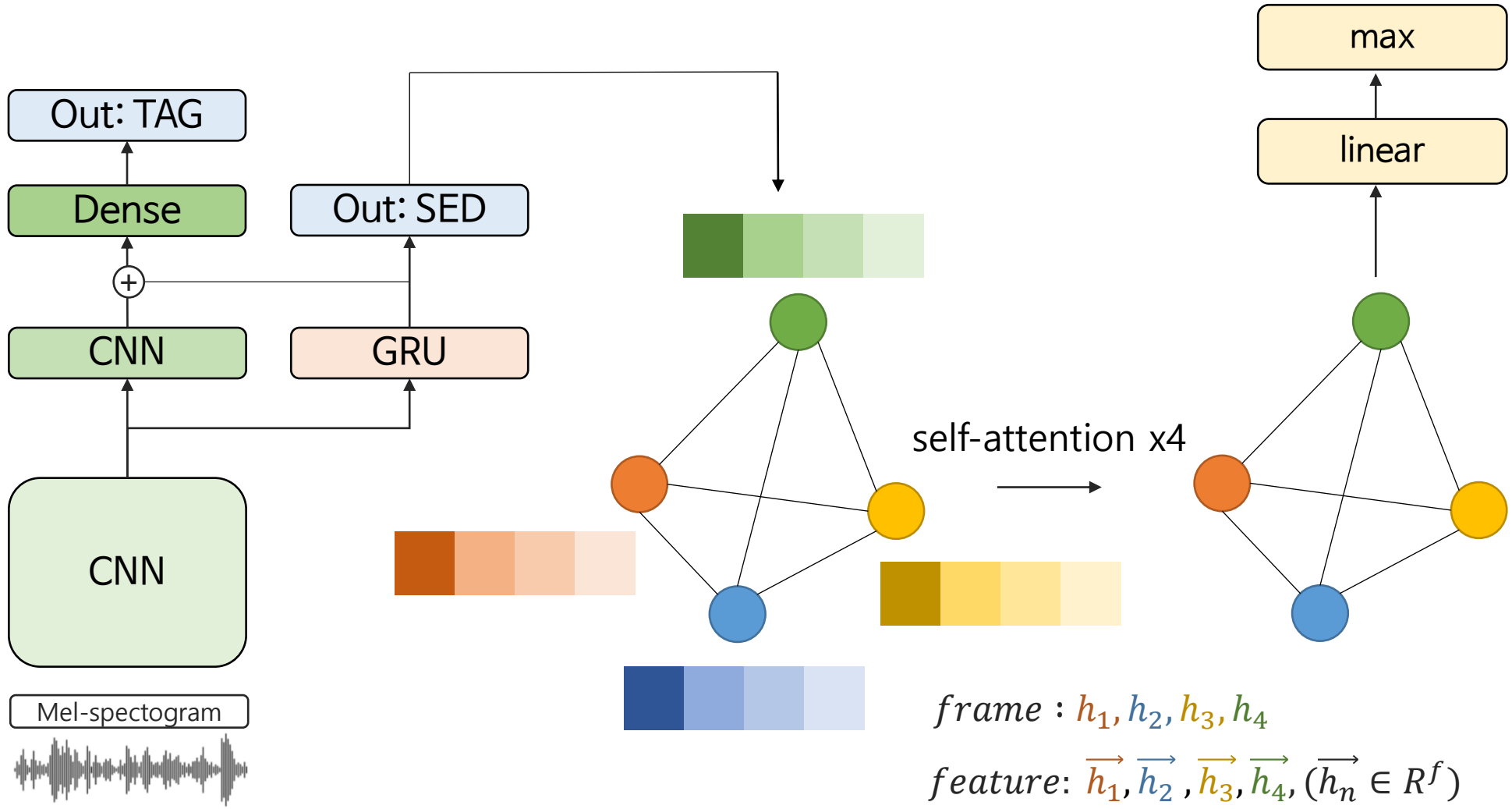


2. DcaseNet + SeResNet

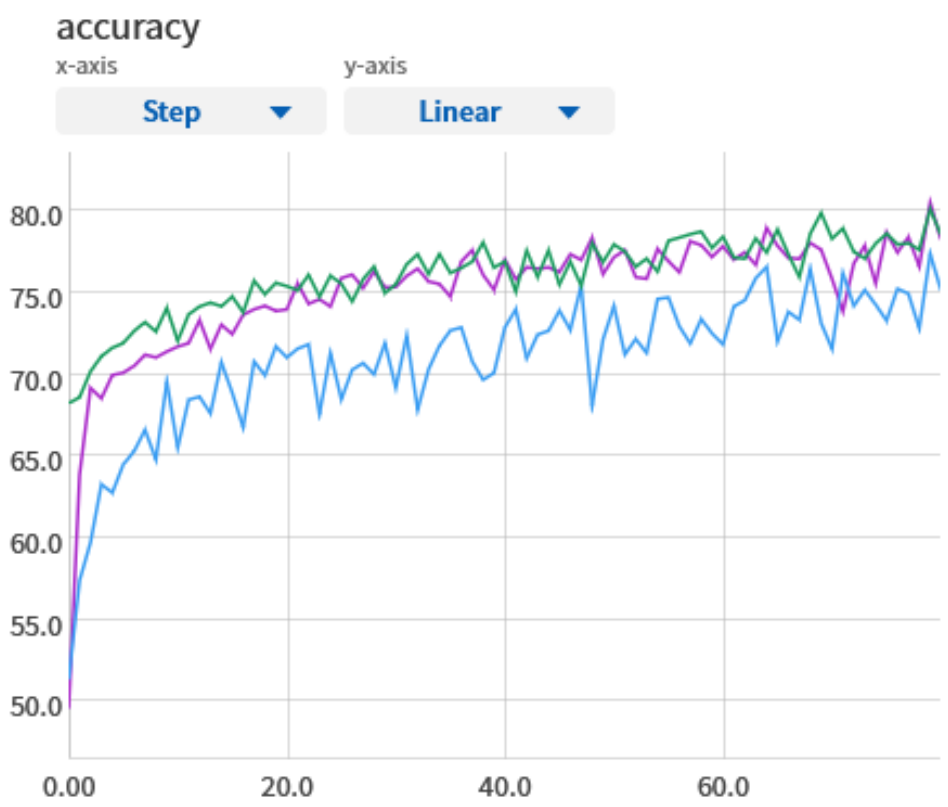


실험 구성 - 2. 모델

3. DCASENet+ GAT

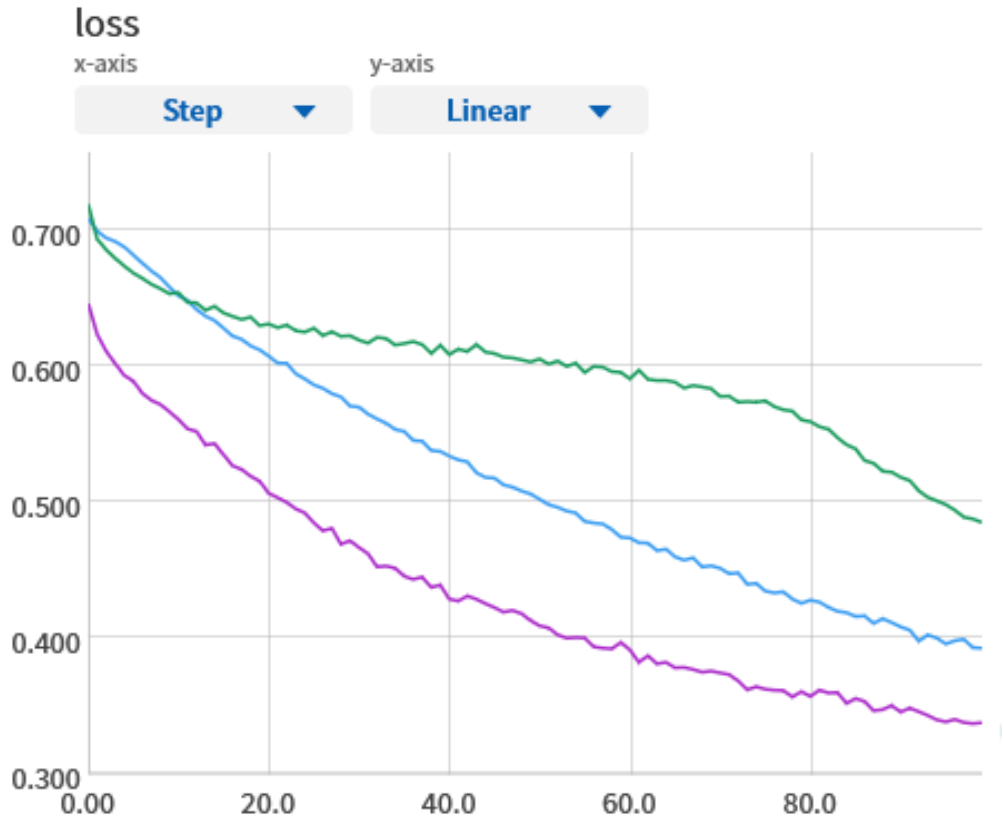


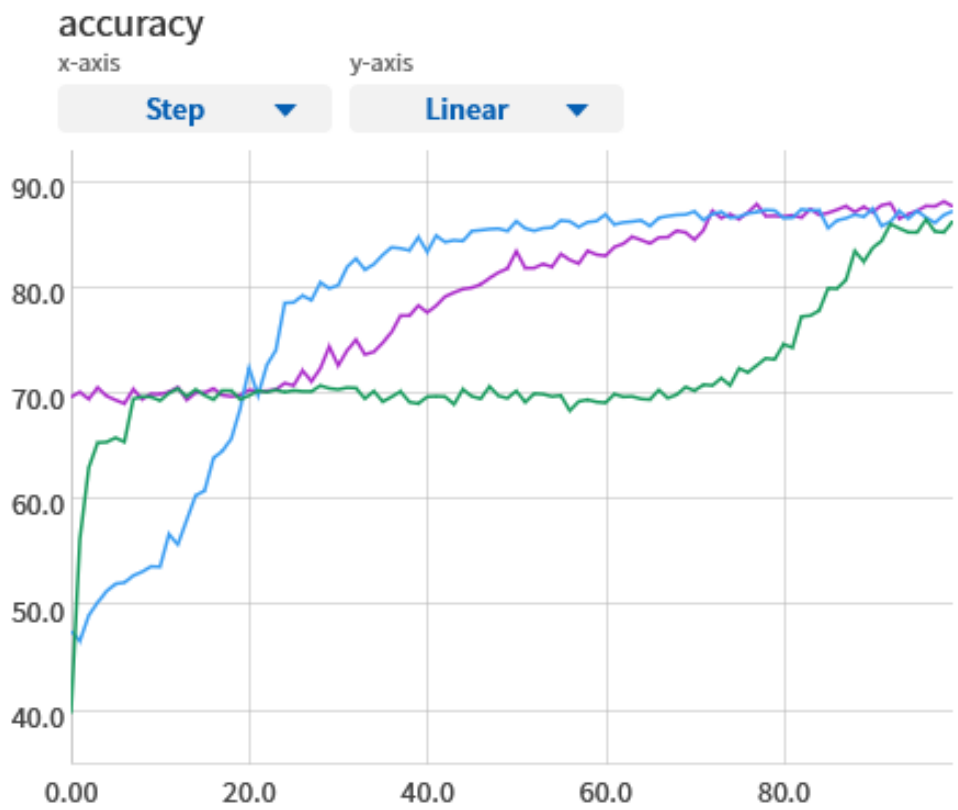
실험 결과



DCASENet + SENet (초록색) 비위험 대 위험 레이블 비율 30%:70%
(파란색) " 50%:50%
(보라색) " 70%:30%

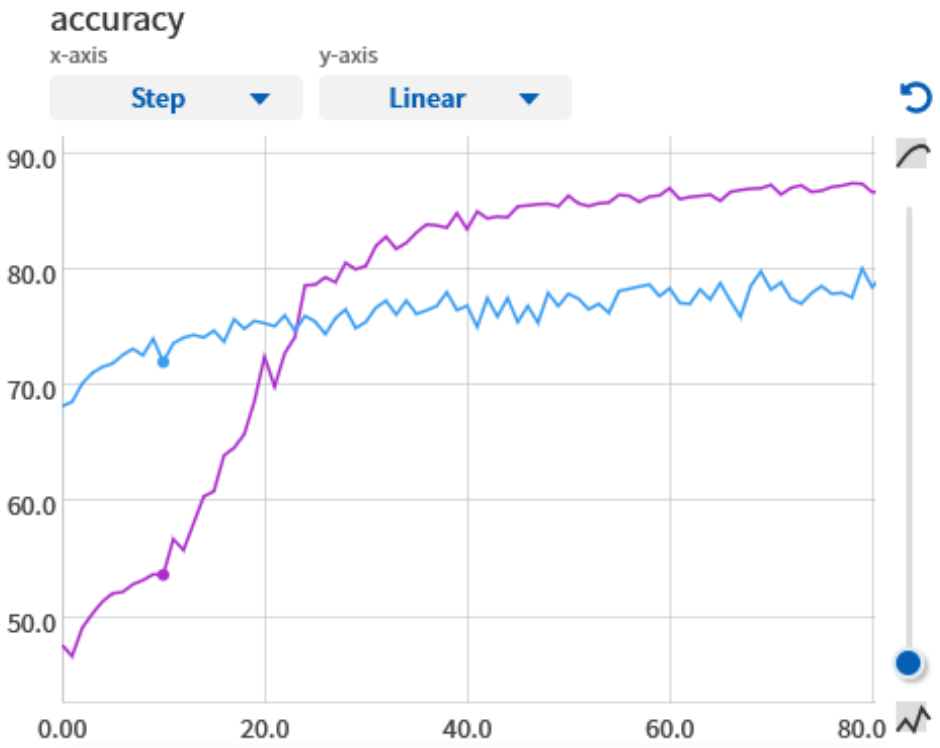
실험 결과





DCASENet + GAT (초록색) 비위험 대 위험 레이블 비율 30%:70%
(파란색) " 50%:50%
(보라색) " 70%:30%

5 결과 분석 및 개선



비위험 대 위험 레이블 비율 50%:50%
(보라색) DCASENet + GAT
(파란색) DCASENet + SeNet

- 결과 분석

- 비위험 대 위험 레이블 비율 50%:50%일 때 가장 높은 정확도 기록
- 실험에서 사용한 모델 중 GAT가 가장 높은 정확도를 기록

- 문제점 및 개선

- Sound Event Detection 에서 Multi-label Classification 대신 Binary Classification 으로 낮은 정확도
- GAT에서 노드 feature로 음성 특징 대신 Event 도메인으로 차원 변환 후 노드 간의 관계 학습 시 성능 개선 기대

Reference

[1] Jung, J. W., Shim, H. J., Kim, J. H., & Yu, H. J. (2020). DCASENET: A joint pre-trained deep neural network for detecting and classifying acoustic scenes and events. *arXiv preprint arXiv:2009.09642*

[2] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[3] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

[4] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

[5] Naranjo-Alcazar, J., Perez-Castanos, S., Zuccarello, P., & Cobos, M. (2020). *TASK 1 DCASE 2020: ASC WITH MISMATCH DEVICES AND REDUCED SIZE MODEL USING RESIDUAL SQUEEZE-EXCITATION CNNS*. DCASE2020 Challenge, Tech. Rep.

[6] Liu, S., Wu, H., Lee, H. Y., & Meng, H. (2019, December). Adversarial attacks on spoofing countermeasures of automatic speaker verification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 312-319). IEEE.

[7] Naranjo-Alcazar, J., Perez-Castanos, S., Ferrandis, J., Zuccarello, P., Cobos, M., & Visualfy, B. TASK 3 DCASE 2020: SOUND EVENT LOCALIZATION AND DETECTION USING RESIDUAL SQUEEZE-EXCITATION CNNS.

[8] Li, L., Gan, Z., Cheng, Y., & Liu, J. (2019). Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 10313-10322).