[소프트웨어학과]

Signal data analysis based on domain generalization

[결과보고서]

Version 1.0
2023 . 06 . 21

학번: 201821505
이름: 최동훈
지도교수: 이 슬

목차

[그림 목차]

[그래프 목차]

요약

Patient covariate can lead to bias and generalization problems if not considered in healthcare applications. This study aims to learn and apply the concept of Domain Generalization to an individualized perspective by using ManyDG (Many-domain Generalization) as the base model structure. In ManyDG, each patient is set as a domain, and domain invariant label is predicted using feature extraction, domain encoding, and orthogonal projection.

The study aims to find the most suitable base learning model for ManyDG, combine it, and measure the accuracy with public medical data to confirm the suitability of the model. Furthermore, Wav2vec2.0 is selected as the most suitable base learning model for better feature extractor. This research aims to create a new model by combining ManyDG with the feature extractor of the wav2vec2.0 model to improve accuracy when dealing with bio signal data such as EEG dataset.

1 개요

1.1 연구의 배경과 목표

기계 학습에서 domain 은 일반적으로 다양한 조건, 환경 또는 특성을 가진 데이터 그룹을 나타낸다. Supervised learning 과 달리 실제 응용에서는 새로운 domain 에서 모델을 평가하거나 배포해야 할 수 있다. 이에 Domain generalization(도메인 일반화)은 모델 훈련 중에 보이지 않는 domain 에서 잘 수행하기 위한 목적으로 여러 source domain 에서 훈련하는 기계 학습 설정을 말한다. 본 연구의 목표는 다양한 domain 또는 분포 변화에 걸쳐 일반화할 수 있는 모델을 개발하여 훈련 데이터와는 다른 domain 에서도 좋은 성능을 발휘할 수 있는 모델을 구축하는 것이다.

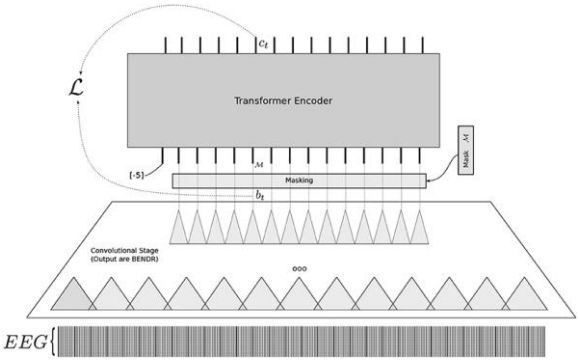
2 관련 선행 연구 조사

2.1 선행 연구

2.1.1 선행 연구 제목

BENDR: USING TRANSFORMERS AND A CONTRASTIVE SELF-SUPERVISED LEARNING TASK TO LEARN FROM MASSIVE AMOUNTS OF EEG DATA (D. Kostas.,2021)

2.1.2 선행 연구 내용



<그림 1 - BENDR Model structure>

Wav2vec2.0 모델의 원래 목표는 Speech recognition(음성 인식) 이지만, 저자들은 이를 EEG(electroencephalogram, 뇌전도) Dataset 을 분류하는 데 적용한다. 이 접근 방식은 방대한 양의 데이터를 수집하고 self-supervised learning 을 통해 raw data signal 데이터의 압축된 표현을 학습할 수 있어 수면 단계 분류를 비롯한 다양한 downstream BCI 및 EEG 분류 작업에 맞게 finetuning 할 수 있다. 이는 수면 단계 분류를 위한 task-specific self-supervision 보다 성능이 뛰어나다.

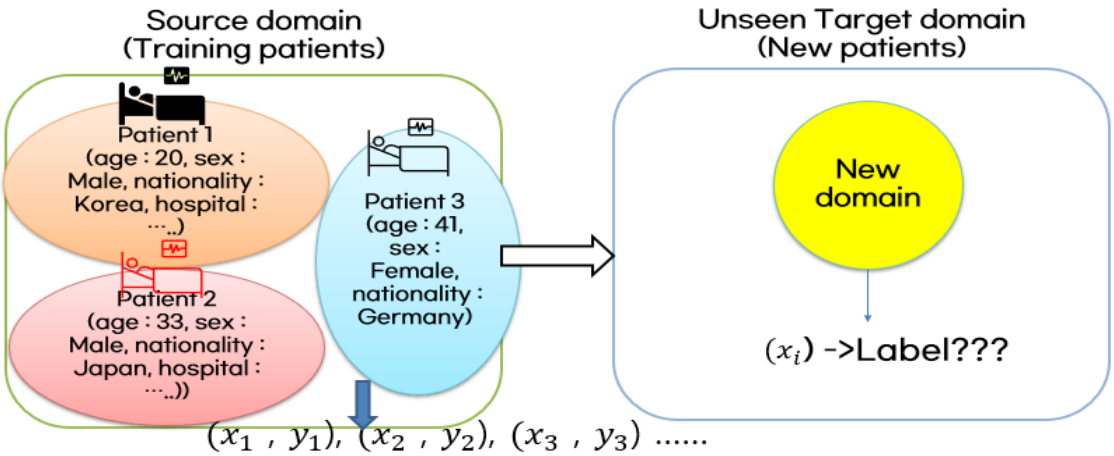
2.1.3 선행 연구와 본 연구간의 차이점

본 연구도 위와 마찬가지로 수면 단계 분류를 위한 작업을 수행한다. 하지만 EEG dataset 은 수집된 환경, 장비나 주체의 차이로 인해 domain 간의 variance 가 크다. 따라서 이를 극복하기 위해 domain generalization 을 활용할 수 있다. Domain generalization 은 다양한 domain 에서 학습된 모델을 새로운 domain 에서 일반화하는 기술이다. 이를 sleep stage classification 에 적용하기 위해서는 다양한 EEG dataset 을 다른 domain 에서 수집하고, 이러한 데이터를 사용하여 모델을 학습한다. 이를 통해 sleep signal 의 특징을 학습하여 새로운 domain 에서도 높은 분류 성능을 보여 일반화할 수 있는 모델을 생성할 수 있다. 추가로 위의 BENDR 에서 제안한 Wav2vec2.0 을 통해 feature vector 를 추출하는 방식을 본 연구의 EEG data 에 활용한다.

3 연구 결과

3.1 문제 정의

3.1.1 문제



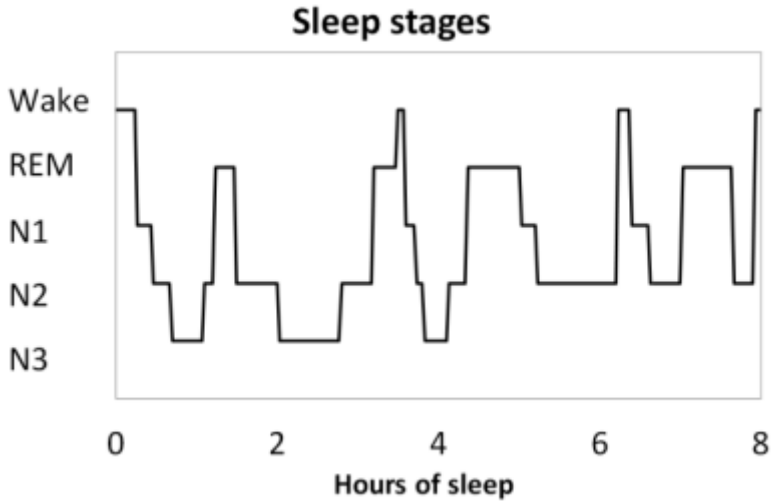
<그림 2 - Patient covariate problem>

임상시험 등의 healthcare applications 에서 환자 데이터는 일반적으로 다양한 성별, 국적, 연령대의 환자가 포함되며, 많은 경우에 각각 남자와 여자, 아시아인과 유럽인, 고령 환자와 젊은 환자 등을 대상으로 다르게 반응할 수 있다. Patient covariate 는 각 환자의 특성 및 데이터 수집 환경과 관련된 고유한 정보이며, 기존의 모델들이 이러한 환자들의 고유한 이질성을 고려하지 않고 학습하면 편향을 유발하고 일반화에 문제가 생길 수 있다.

3.1.2 Dataset

본 연구에서 수면 signal 데이터로 활용하는 Sleep-EDF 데이터는 주로 다중 채널 형태로 제공되며, 각 채널은 특정 생체 신호를 나타낸다. 총 78 명의 수면 기록이 있으며, 이는 415,089 개의 30 초 길이 sample 로 분할될 수 있다. 이 기록들은 100Hz 로 sampling 되었으며

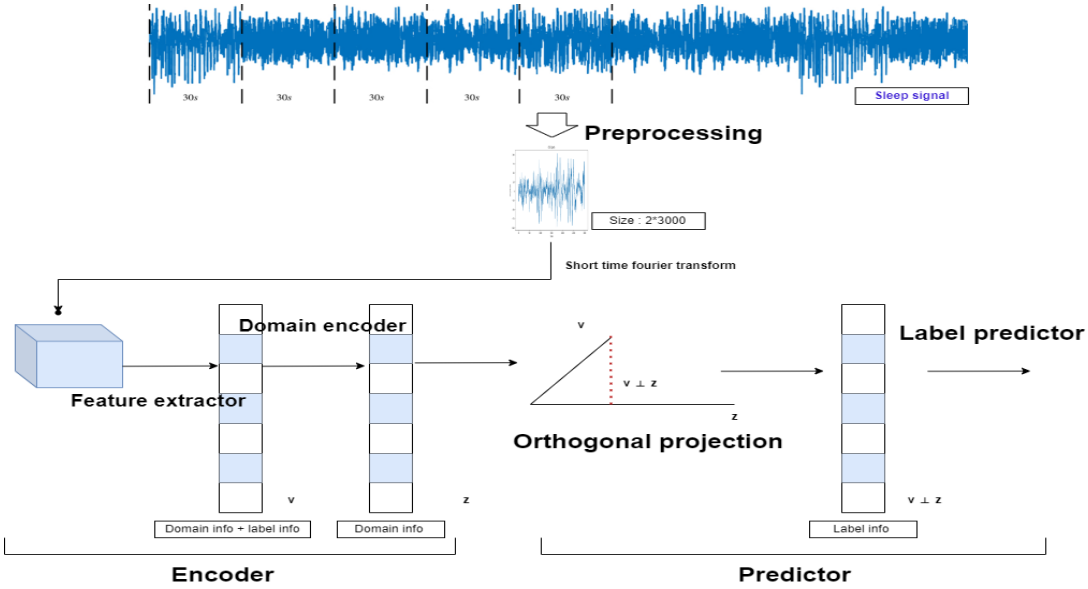
모델의 입력으로 Fpz-Cz/Pz-Oz 채널을 추출한다. 각각의 sample 은 깨어 있는 상태 (Wake), 빠른 안구 운동 (REM) 및 세 가지 수면 상태 (N1, N2, N3) 중 하나로 분류된다.



<그림 3 - 수면 상태의 단계적 분류>

3.1.3 연구 내용

본 연구에서는 수면 단계 예측 분류 모델의 일반화 가능성을 높이기 위해 환자 각각을 하나의 domain 으로 설정하여 학습할 수 있는 domain 의 양을 늘린 ManyDG (Many-domain Generalization) 방법을 사용한다. 이는 소수의 domain 을 가정하는 대부분의 기존 domain 일반화 방법과 비교할 때 독특한 설정이다.



<그림 4 - ManyDG Model>

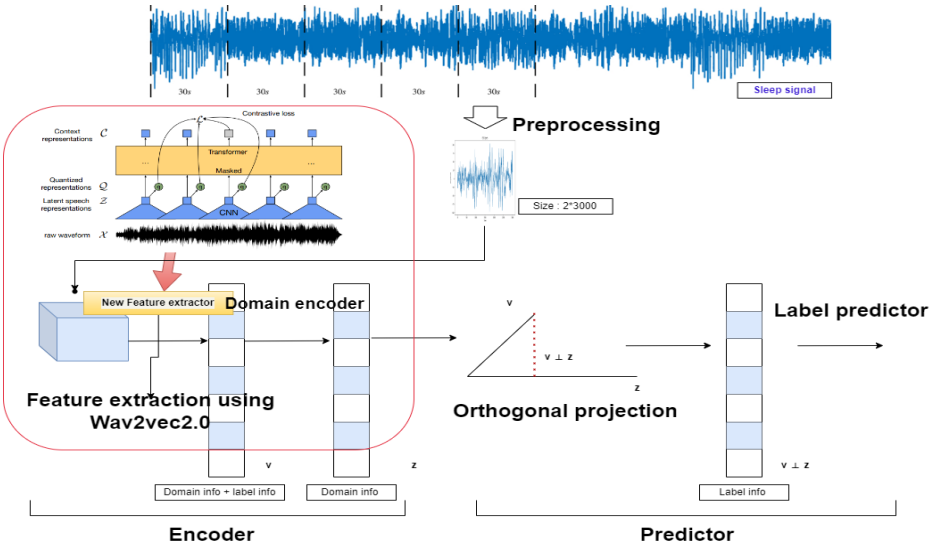
수면 signal 데이터를 전처리하고 sampling 한 뒤에 feature extraction 에서는 각 sample 에 대해서 feature vector v 를 추출한다. Vector v 는 domain 정보와 label 정보를 모두 갖고 있어서 encoder 를 통해 domain 정보가 담긴 latent presentation vector z 를 만든다. 이후 v 벡터와 z

벡터를 활용하여 orthogonal projection 을 거쳐 domain invariant 한 label 과 관련된 vector 만 추출한다. 이 vector 를 통해 수면 단계의 분류에 대한 예측을 수행할 수 있다.

3.2 제안하는 기법 혹은 소프트웨어 구조

3.2.1 ManyDG model + Wav2vec2.0

Wav2vec2.0 은 음성 데이터만으로 강력한 음성 표현을 학습하는 self-supervised 학습 방식이다. 기존 domain generalization 모델의 feature extractor 를 wav2vec2.0 의 것으로 바꿔서 새로운 모델을 만든 뒤 수면 데이터를 대상으로 학습하며 기존 모델과 비교했을 때 성능의 개선 여부를 확인하고자 한다. 비록 수면 데이터와 음성 데이터는 modality 가 다르지만 같은 signal 데이터이기 때문에 wav2vec2.0 의 Transformer 모델로 인한 성능 개선을 기대해 볼 수 있다.



<그림 5 – ManyDG model + Wav2vec2.0>

4 성능 분석

4.1 성능 분석 환경

기존 ManyDG 모델을 통하여 학습을 진행한 것을 1 차 실험, feature extractor 를 Wav2vec2.0 의 것으로 바꿔 학습을 진행한 것을 2 차 실험이라 명명한다. 이를 통한 학습의 결과를 살펴보면 Loss function 의 값이 점점 작아지게 된다. 이는 모델이 예측한 값이 실제 값에 더 가까워졌다는 것을 의미한다.

각 목적 함수는 모델 학습 프로세스에서 특정 목적을 수행하며 Loss_total(총 손실함수)은 이러한 목적 함수의 합이다. 우선, Reconstruction loss 는 모델이 입력 데이터를 정확하게 재구성할 수 있도록 보장한다. Mmd loss 는 서로 다른 domain 간의 분포 격차를 줄이는 데 도움이 된다. 반면 Similarity_loss 는 같은 환자의 두 latent factor(잠재 요인)의 유사성을 적용한다. 또한 Cross - entropy loss 를 최소화함으로써 예측을 개선하고 실제 라벨과 일치하도록 학습한다. 이렇게 함으로써 손실함수는 모델을 조정하여 예측된 라벨과 실제 라벨 사이의 차이를 최소화하는

방향으로 나아간다. 결과적으로 patient covariate 를 제거하여 여러 healthcare applications 에서 일반화 성능을 개선할 수 있다.

4.2 성능 분석 결과

4.2.1 1 차 실험

4.2.1.1 1 차 실험 조건

[실험을 위한 hyperparameter]

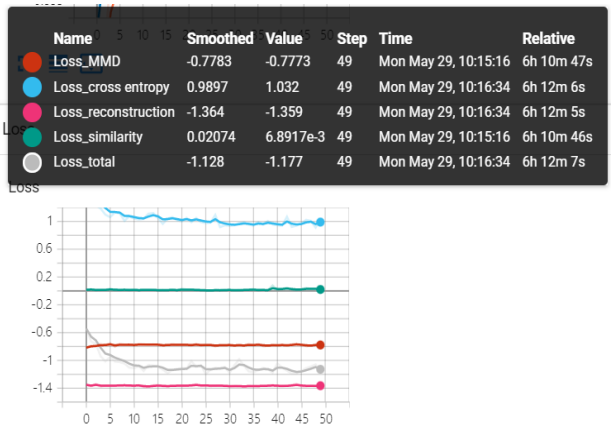
learning time: about 6 hours 13 minutes

Batch_size : 256

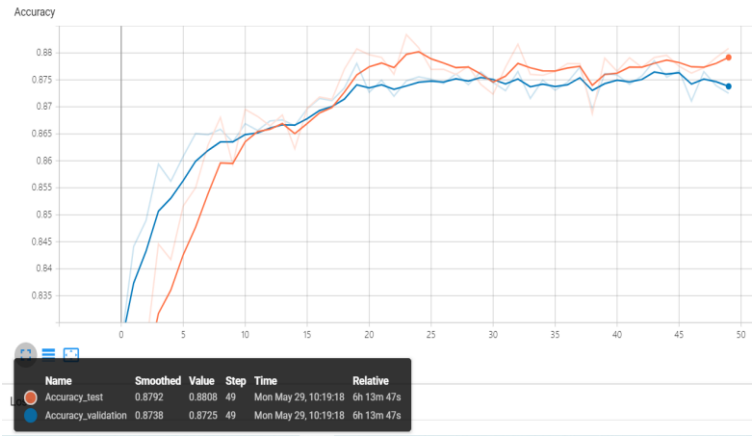
Epochs : 50

5 × 10−4 as the larning rate with Adam optimizer, and 1 × 10−5 as the weight decay

4.2.1.2 1 차 실험 결과



<그래프 1 - 1 차 실험 loss 그래프>



<그래프 2 - 1 차 실험 Accuracy 그래프>

test accuracy 가 epoch 50 에서 0.8789 까지 오른다.

4.2.2 2 차 실험

4.2.2.1 2 차 실험 조건

[실험을 위한 hyperparameter]

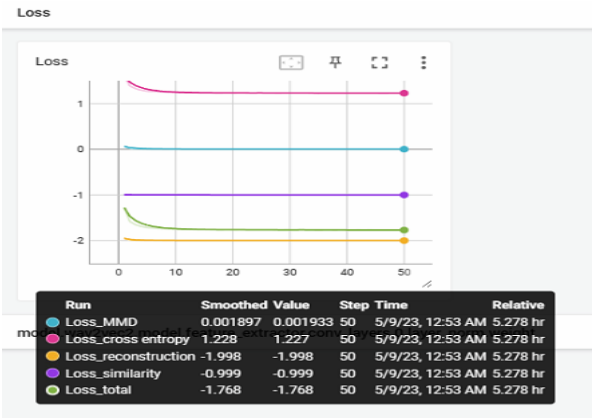
learning time: about 5 hours 16 minutes

Batch_size : 128

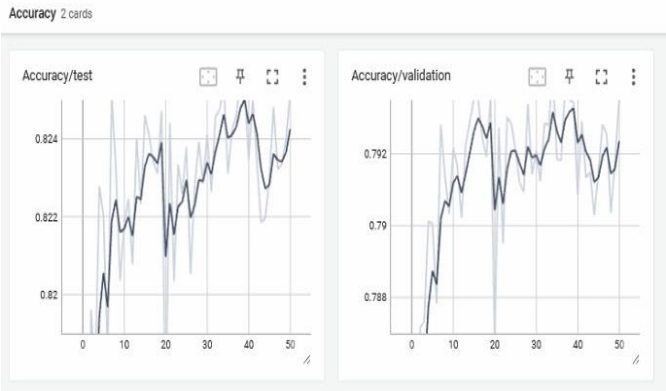
Epochs : 50

5 × 10−4 as the larning rate with Adam optimizer, and 1 × 10−5 as the weight decay

4.2.2.1 2 차 실험 결과



<그래프 3 - 2 차 실험 loss 그래프>



<그래프 4 - 2 차 실험 Accuracy 그래프>

test accuracy 가 epoch 50 에서 0.8243 까지 오른다.

5 결론

본 연구의 목표는 다양한 domain 또는 분포 변화에 걸쳐 일반화할 수 있는 모델을 개발하여 훈련 데이터와는 다른 domain 에서도 좋은 성능을 발휘할 수 있는 모델을 구축하는 것이다. 기존 domain generalization 방법과 비교했을 때 ManyDG 모델의 정확도는 개선되었으며 loss 또한 일정하게 하락하는 것으로 보아 정상적으로 학습이 된다. Task 간의 융합이 일어난 2 차 실험에서도 학습이 정상적으로 진행된다. 서로 다른 modality 의 task 여서 기대를 많이 하지

않았지만 구조를 많이 바꾸지 않은 채 일부 레이어만 freezing 하여 학습을 진행하였고 정확도의 개선이 일부 이뤄진 것을 확인할 수 있다. 이를 통해 서로 다른 task 간의 결합을 통해 기존 모델보다 성능이 개선된 새로운 모델을 만들 수 있을 거라는 가능성을 확인하였다.

하지만 Wav2vec2.0 모델이 음성 기반 인식 모델이기 때문에 data 의 modality 차이에 의해 정확도가 올라가는 것이 한계가 있음을 추론할 수 있다. 따라서 Wav2vec2.0 모델의 feature extractor 를 sleep signal 에 맞게 수정하고 이를 pretraining 할 계획이다. 즉, Sleep signal 을 통해 self-supervised learning 을 실험하여 다른 dataset 을 대상으로 downstream task 를 진행하여 정확도를 파악할 계획이다. 이와 같은 과정을 통해 Pretrain 된 wav2vec2.0 의 feature extractor 로 새로운 모델을 만들어 finetuning 을 진행하여 성능의 개선을 끌어내고자 한다.

참고자료

[1] C. Yang, W. M. Brandon, and J. Sun, "ManyDG: Many-domain Generalization for Healthcare Applications," Jan. 2023, doi: https://doi.org/10.48550/arxiv.2301.08834.

[2] Alexei Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," vol. 33, pp. 12449–12460, Jun. 2020.

[3] D. Kostas, S. Aroca-Ouellette, and F. Rudzicz, "BENDR: Using Transformers and a Contrastive Self-Supervised Learning Task to Learn From Massive Amounts of EEG Data," Frontiers in Human Neuroscience, vol. 15, Jun. 2021, doi: https://doi.org/10.3389/fnhum.2021.653659.

[4] M. Wu, Y. Lu, W. Yang, and S. Y. Wong, "A Study on Arrhythmia via ECG Signal Classification Using the Convolutional Neural Network," Frontiers in Computational Neuroscience, vol. 14, Jan. 2021, doi: https://doi.org/10.3389/fncom.2020.564015.

[5] S. Mousavi, F. Afghah, and U. R. Acharya, "SleepEEGNet: Automated sleep stage scoring with sequence to sequence deep learning approach," PLOS ONE, vol. 14, no. 5, p. e0216456, May 2019, doi: https://doi.org/10.1371/journal.pone.0216456.

[6] S. Alvarez, "From covariates to confounding factors: the danger of having too many covariates," Cognivia, Oct. 20, 2020. https://cognivia.com/from-covariates-to-confounding-factors-the-danger-of-having-too-many-covariates/ (accessed Jun. 21, 2023).

[7] "Sleep-EDF Database Expanded v1.0.0," www.physionet.org. https://www.physionet.org/content/sleep-edfx/1.0.0/ (accessed Jun. 21, 2023).

[8] "[Vision AI Workshop] 카이스트 주재걸 교수 - Domain Generalization and Out-of-Class Detection in ...," www.youtube.com. https://www.youtube.com/watch?v=GYp-Hfvf_T0 (accessed Jun. 21, 2023).

[9] "Supervised learning". Available: https://www.diegocalvo.es/en/supervised-learning/ (accessed Jun. 21, 2023).

[10] "Visual Studio Tools 다운로드 - Windows, Mac, Linux 용 무료 설치," Visual Studio. https://visualstudio.microsoft.com/ko/downloads/ (accessed April. 21, 2023).

[11] "pip install Microsoft Visual C++ Build Tools 에러". Available: https://ddbobd.tistory.com/entry/python-Microsoft-Visual-C-Build-Tools-%EC%97%90%EB%9F%AC (accessed Jun. 7, 2023).

[12] "오류 코드 1 로 실패한 'Python Setup.py egg_info'를 해결하는 방법". Available: https://www.easeus.co.kr/data-recovery/python-setup-py-egg-info-failed-with-error-code-1.html (accessed Jun. 5, 2023).

[13] "RuntimeError: Distributed package doesn't have NCCL built in". Available: https://discuss.pytorch.org/t/runtimeerror-distributed-package-doesnt-have-nccl-built-in/176744/10 (accessed Jun. 1, 2023).

[14] "torch.utils.tensorboard — PyTorch 2.0 documentation". Available: https://pytorch.org/docs/stable/tensorboard.html (accessed May. 21, 2023).

[15] "Fine-Tune Wav2Vec2 for English ASR in Hugging Face with Transformers". Available: https://huggingface.co/blog/fine-tune-wav2vec2-english (accessed Jun. 11, 2023).

[16] "ICLR'23 ManyDG Paper", Jun. 21, 2023. Available: https://github.com/ycq091044/ManyDG (accessed Jun. 18, 2023).