A Report of System Design (Final)

A Universal EEG Encoder for Image & Auditory Reconstruction

Graduate Project 1 (3183) - Karpjoo Jeong
Team 7, EEmaGe – Advisor : Eun Yi Kim
Computer Science and Engineering 201811218 Hyun Woo Lee
Computer Science and Engineering 201811182 Wonjun Park
Computer Science and Engineering 201911193 Esan Woo

Abstract

이 보고서는 EEG (Electroencephalography) Brain Signals로부터 인간의 시각 이미지와 청각 소리를 재현하는 연구의 시스템 설계서입니다. EEG data는 Brain-Computer Interface (BCI) 분야에서 각광받고 있는 비용 효율적인 데이터 중 하나입니다. 기존의 연구들은 EEG-Image datasets이나 EEG-Audio datasets 등을 가지고, EEG 데이터를 녹음할 때 보여진 시각(이미지) 데이터나 청각(소리) 데이터를 재현해내는데 어느 정도 성과를 보이고 있습니다. 하지만 지도 학습 방법의 사용과 통합된 인코더의 부재라는 한계점이 존재합니다. 본 연구는 다음의 두 가지 가설을 제시합니다: *1) 인간 체계를 모방하기 위해서는 Supervision 이 아닌 다른 방법이 필요하다. 2) 이미지와 소리에 대한 인공지능 모델 학습이 상관성을 지닌다.* 이를 토대로, 연구는 궁극적으로 (Autoencoder 등의) Unsupervised Learning을 적용한 A Universal EEG Encoder의 제안을 목표로 합니다. 본 보고서의 본론에서는 이 Universal Encoder의 설계와, 현재 초점을 맞추고 있는 부분을 중심으로 한 설계를 소개합니다.

I. Introduction

BCI (Brain-Computer Interface)는 뇌와 컴퓨터를 연결하여 뇌의 신호를 분석하고 이를 통해 컴퓨터나 기계를 제어하는 기술입니다. BCI 연구는 뇌의 신호를 분석하여 뇌의 기능을 이해하고, 뇌의 장애를 극복하거나 뇌의 기능을 향상시키는 데에 활용하고자 하는 목적으로 행해져 왔습니다. 연구에서 뇌 신호를 분석하는 방법은 다양합니다. fMRI (functional Magnetic Resonance Imaging), BOLD (Blood Oxygen Level-Dependent), ECoG (Electrocorticography), MEG (Magnetoencephalography) 등이 있습니다. 특히 fMRI로 읽은 뇌 신호는 상대적으로 높은 재현율을 보여왔습니다 [1][2][3]. 하지만 fMRI 장비는 장비 사용도에 제한성이 높다는 단점이 존재합니다. EEG (Electroencephalography)는 뇌의 표면에서 발생하는 전기적인 신호를 측정하는 방법으로, 비용이 저렴하다는 장점이 있습니다. 따라서 본 프로젝트에서는 EEG 데이터를 이용하여 뇌 신호를 분석합니다.

인공지능, 그 중 기계학습은 일반적으로 네 가지의 카테고리로 분류됩니다. Supervised Learning (지도학습), Unsupervised Learning (비지도학습), Semi-supervised Learning (반지도학습), 그리고 Reinforcement Learning (강화학습) 입니다. 지도학습은 모델 학습 과정에서 주어진 Input에 대해 답을 알려주어 학습시키는 것이고, 비지도학습은 답을 알려주지 않고 학습시키는 방법입니다. 현재 뛰어난 성과를 보이고 있는 인공지능 모델들은 대부분 지도학습에 속합니다. 하지만 연속성이 있는 시계열 데이터를 분석하는 데에는 아직까지 한계가 존재합니다. EEG 데이터는 비선형적이며 시계열 데이터로 구분되며, 이러한 데이터들이 가지는 어려움을 공유합니다. 기존의 연구는 적합한 데이터 수집 실험 설계와 인공지능 모델 설계를 하여 이를 타파하고자 했으며, 괄목할만한 수준의 성능을 보이고 있습니다 [4][5].

하지만 현재까지의 연구는 인코딩 방법에 지도학습을 사용했다는 한계점이 존재합니다. 인간의 시각 체계가 형성되는 데에는 Supervision이 필요하지 않습니다 *(가설 1).* 본 연구에서는 Autoencoder 등의 비지도학습을 사용해 EEG Encoder를 학습함으로써, 이 명제를 증명하고자

합니다. 한편 이미지를 생성해내는데 사용하는 Decoder에서는 확률 기반 생성 모델은 배제합니다. 실제 경험했던 것을 똑같이 재현해내는데에 초점을 맞추기 위함입니다.

이미지와 소리는 서로 다른 도메인에 속하는 데이터입니다. 하지만 두 데이터는 값의 연속성이 있다는 점에서 그 상관성이 있습니다. 실제로 [6][7]과 같이, 이 두 가지 데이터를 함께 분석하려는 시도들이 있었습니다. 본 연구 또한 이미지와 소리가 상관성을 가져, 통합된 인공지능 모델을 이용한 분석이 가능하다는 가설 *(가설 2)* 을 뒷받침하고자 합니다.

ICASSP 2024: A EEG Auditory Challenge [8]는 EEG와 소리 데이터 간의 상관관계를 찾는 공개 챌린지입니다. 이 챌린지의 Deadline이 2023년 12월 28일이기 때문에, 본 연구에서는 현재 청각 데이터의 재현에 우선적으로 실험을 진행하고 있습니다. 현재 이 챌린지를 해결하기 위해 세 가지 접근을 하고 있습니다. 데이터 분석과 전처리, 시각 재생성 인코더의 활용, 신뢰도 높은 다른 청각 재현 모델 사용. 각 접근에 대해서는 *II. Related Works*와 *III. System Overview* 에서 상세히 다룹니다.

결론적으로 본 연구는 EEG 데이터로부터 시각과 청각을 모두 재현할 수 있는 Universal EEG Encoder의 제안을 목표로 합니다. 이 과정에서 Unsupervised Learning을 이용한 새로운 EEG 데이터에 대한 프레임워크를 제안하여, Neuroscience와 Artificial Intelligence (AI)의 Multidisciplinary 영역에 기여할 수 있을 것입니다. 이를 통해 뇌의 기능을 이해하고, 뇌의 장애를 극복하거나 뇌 기능을 향상시키는 데에 활용할 수 있는 기술의 근간을 마련할 수 있을 것이라 기대합니다.

II. Related Works

a. Visual Image Reconstruction

EEG 뇌파로부터 시각 이미지를 재생성하는 연구는 일정 수준의 이미지 생성에 성공하여 왔습니다. Khare, Sanchita, et al. [4]은 Decoder의 성능 개선에 집중한 NeuroVision 프레임워크를 제안하여, 보다 선명한 이미지를 얻으려 하였습니다. Progressive Growing of GAN [10]을 활용한 Conditional ProGAN을 제안해, 보다 선명한 이미지를 재현해내는데 성공하였습니다. 하지만 Encoder에서는 여전히 LSTM과 GRU 등의 신경망 알고리즘을 이용해 Supervised Learning을 사용했다는 점에서, 본 연구에서 지적하는 *가설 1*의 문제를 해결하지 못했다는 한계가 존재합니다. Ye, Zesheng, et al. [5]은 *가설 1*에 대한 인식을 공유하며, Supervised Method에서 탈피하고자 한 연구입니다. 이 연구에서는 "Is the class-specific reconstruction with GANs the only way to restore visual stimuli from EEG Signals?"라는 질문을 던지며 문제점을 제기하였고, Self-supervised Method와 Multimodality를 채택해 시각 재생성 파이프라인을 만들었습니다. 그럼에도 불구하고 인코딩 방법에 multimodality를 통해 visual cue를 제공했다는 점에서 그 한계가 있습니다.

한편 확률 기반 생성 모델을 이용한 연구도 이용 가능한 데이터셋을 다양하게 만들었다는 점에서 본 연구에서 사용할 수 있다는 의의가 있습니다. Singh, Prajwal, et al. [11]은 EEG2IMAGE라는 프레임워크를 제안해, 작은 EEG 데이터셋으로부터 여러 다른 이미지를

생성해낼 수 있도록 했습니다. EEG 데이터 수집에 어려움이 있다는 점을 고려했을 때, 이 프레임워크는 더 다양한 데이터셋의 사용을 가능하게 만들어줍니다.

b. Auditory Stimuli Reconstruction

EEG로부터 청각 자극을 재생성하는 연구 또한 여럿 이루어져왔습니다. SparrKULee [12]는 85명의 참가자를 대상으로, 11개로 라벨링된 약 90-150분의 Audio Stimuli를 주어 만든, 총 168시간 규모의 데이터셋입니다. Thornton [13]은 50개의 모델을 앙상블 기법을 이용해 학습해, EEG와 매칭되는 Audio Stimuli를 2개의 선택지 중에서 찾아내는 문제에 대해 약 82%의 정확도를 보였습니다. Piao, Zhenyu, et al. [14]는 Transformer와 Wav2Vec의 Local Optimization 개념을 이용하여, 참가자들의 unique한 EEG 데이터 특성을 반영하여 청각 자극을 재생성하였습니다. 성능 평가는 Pearson Correlation을 이용하였으며, 약 0.159의 성능을 보였습니다.

본 연구에서는 (1) 5개의 선택지에서 EEG와 매칭되는 Audio Stimuli를 찾는 문제와 (2) EEG로부터 청각 자극 (Mel-spectogram)을 재생성하는 문제를 해결하기 위해, 다음의 두 가지 접근을 하고 있습니다. 해당 접근들과 관련된 연구들은 다음과 같습니다.

시각 재현 이미지 인코더 활용 *가설 2*에 의거하여, 이미지 재생성에 사용된 EEG Encoder를 청각 재현에 사용하여도 좋은 성능을 보일 것이라 생각해 관련 실험을 진행하고 있습니다. EEG-ChannelNet은 Palazzo, Simone, et al. [9]의 연구에서 사용한 EEG Encoder입니다. 해당 연구에서는 Siamese Network와 Triplet Loss를 활용하여 (EEG-Image) 간의 페어를 학습합니다. 최종적으로 60% 이상의 정확도를 보인 인코더입니다.

신뢰도가 높은 다른 청각 재현 인코더 활용 Meta AI의 Defossez, Alexandre, et al [15]는 최근 EEG, MEG(Magnetoencephalography, 뇌자도) 둘 다 입력으로 받을 수 있는 청각 재현 모델을 제안했습니다. 본 연구에서는 wav2vec 2.0 모델을 활용한 speech representation 모듈과 Spatial Attention과 CNN을 활용한 Brain(encephalography) 모듈에서 추출된 feature vector들, 그리고 modality가 다른 두 vector들 간의 feature들의 조합을 찾는 것에 효율적인 CLIP(Contrastive Language-Image Pre-training) Loss를 활용했습니다. 이러한 구조를 가진 모델에서 Top-10 accuarcy task(테스트 셋에 있는 음성 segment 중 하나의 segment가 주어지면 모델이 추론한 예측 확률 중 Top 10에 해당 세그먼트가 들어있는 경우 정답인 태스크)의 정확도는 71%까지 보였습니다.

III. System Overview

본 섹션에서는 *II.a. A Universal EEG Encoder*와 *II.b. Data Analysis and Preprocessing, II.c. An Architecture of Auditory Specialized Encoder* 에 대해 소개하고자 합니다. 본 연구에서 궁극적으로 제안하고자 하는 범용 EEG 인코더를 먼저 설명하고, 현재 챌린지에서 집중하여 풀고자 하는 모델을 설명합니다. *II.c*는 이상적인 성능의 *II.a*를 제안하기 위해, EEG와 Audio의 상관성을 분석, 입증하고 효과적인 모델 구조를 찾기 위한 필수적인 단계입니다.

a. A Universal EEG Encoder

현재까지 설계한 구조는 다음과 같습니다.

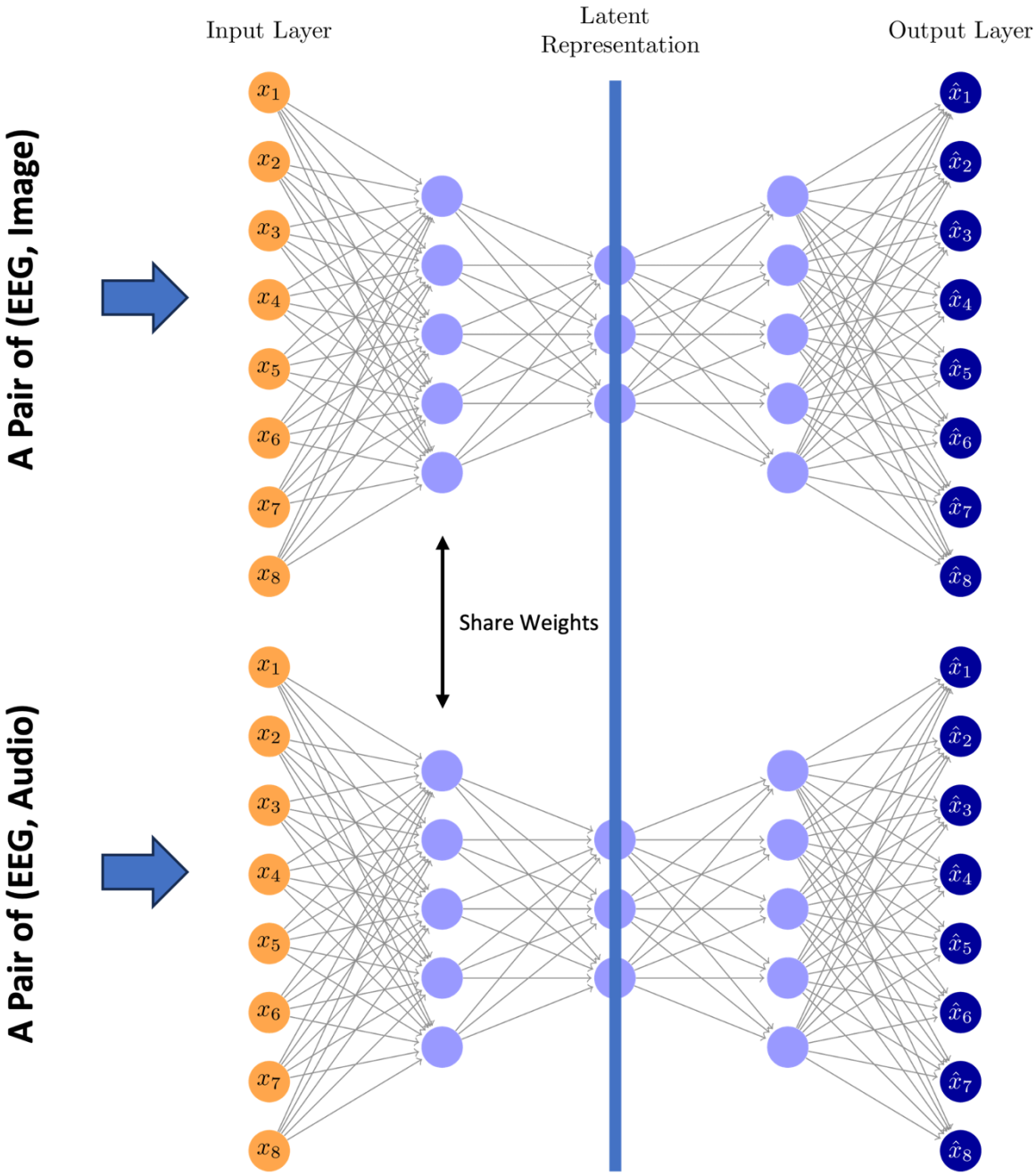

Figure 1. A Model of an Universal Encoder Based on Autoencoder

두 개의 Autoencoder를 이어붙여 (EEG, Image)와 (EEG, Audio)를 함께 Input으로 받습니다. Encoder 부분에서는 두 Autoencoder가 가중치를 공유하도록 합니다. Decoder에서는 각각 Image와 Audio에 대한 loss value를 계산하고, 둘의 가중평균 값을 최종 loss 값으로 합니다.

범용 EEG 인코더가 될 부분은 Autoencoder의 앞부분, 가중치를 공유하는 부분입니다. Image와 Audio에 대해 적절한 Feature Vector를 뽑아내는 것을 목표로 합니다. 이는 EEG로

부터 뇌파의 일정한 규칙을 찾아낼 수 있을 것이란 기대를 수반합니다.

b. Data Analysis and Preprocessing

생체 정보는 같은 것에 대해서도 사람마다 다를 수 있다는 특징을 가집니다. EEG 채널별로, 피실험자별로 데이터를 분석합니다.

EEG Channel별 Correlation

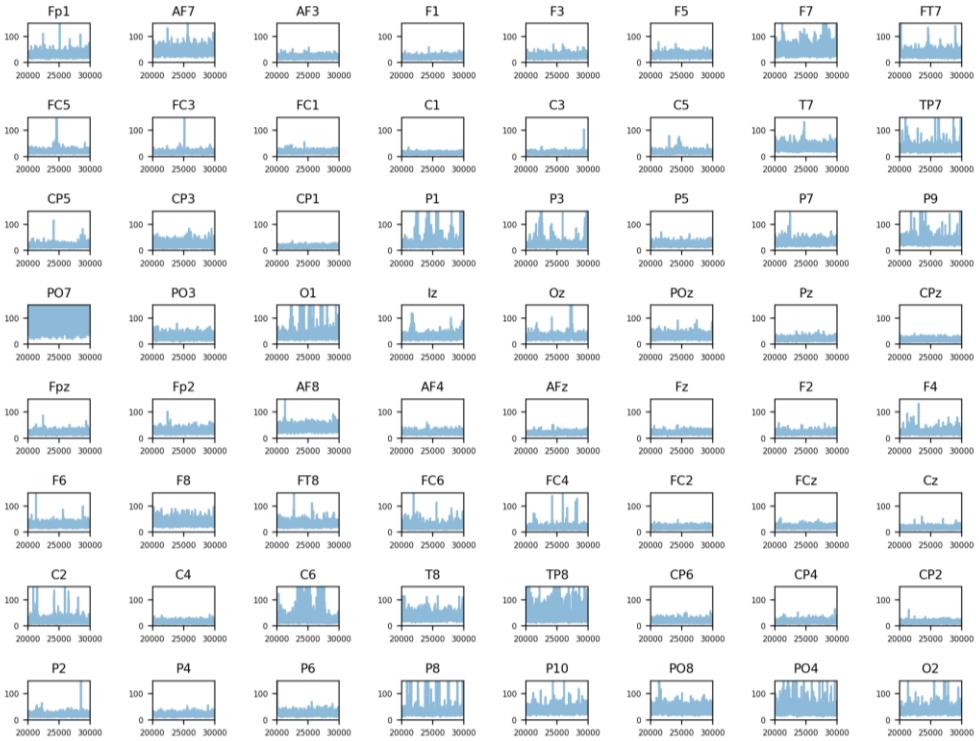

Figure 2. Differences for 64 EEG Channels

위의 그래프는 EEG의 64개 채널에 대해, 피실험자들의 Difference를 찍은 그래프입니다. 하지만 Subject별로 Correlation이 높아 보이는 채널은 관측되지 않았습니다.

피실험자별 Normalization

# epoch	# accuracy	# loss	# val_accuracy	# val_loss
0	0.5609015822410583	1.1048883199691772	0.4721872806549072	1.3222718238830566
1	0.5887477993965149	1.0565221309661865	0.5408531427383423	1.181067943572998
2	0.6025387644767761	1.0259857177734375	0.5456657409667969	1.150052785873413
3	0.6150955557823181	0.9989422559738159	0.5521615743637085	1.1429476737976074
4	0.6198536157608032	0.9847458004951477	0.5545756816864014	1.1418417692184448
5	0.6253008842468262	0.9729302525520325	0.5647258162498474	1.1158109903335571
6	0.627524197101593	0.9666241407394409	0.5677400827407837	1.1226760149002075
7	0.6315232515335083	0.9574078917503357	0.5687322616577148	1.1103869676589966
8	0.6327700614929199	0.953141987323761	0.5702471733093262	1.1042563915252686
9	0.6376044750213623	0.9445868730545044	0.5584514737129211	1.1429566144943237
10	0.6378390789031982	0.9419689774513245	0.5766588449478149	1.0873609781265259
11	0.6365922689437866	0.9448838233947754	0.5708473324775696	1.1049293279647827
12	0.6427754759788513	0.9322583079338074	0.5705904364585876	1.1041871309280396
13	0.6435666680335999	0.9279400706291199	0.5789688229560852	1.1073412895202637
14	0.6445266008377075	0.9266883730888367	0.5747718811035156	1.0954147577285767
15	0.6439569592475891	0.9280425310134888	0.5741872191429138	1.1100633144378662

Table 1. Training-log in a Simple ML Model Without Normalization

# epoch	# accuracy	# loss	# val_accuracy	# val_loss
0	0.5874829292297363	1.0453938245773315	0.4885697066783905	1.293883204460144
1	0.6114495992660522	1.0071587562561035	0.5419560670852661	1.1638498306274414
2	0.6265190243721008	0.9703983068466187	0.5669870376586914	1.1128772497177124
3	0.6334376335144043	0.9555190801620483	0.5656737089157104	1.127506136894226
4	0.6403657793998718	0.9385828971862793	0.558464765548706	1.1494619846343994
5	0.6450394988059998	0.9263946413993835	0.5891721248626709	1.0701382160186768
6	0.6463465690612793	0.9239087700843811	0.5862553119659424	1.0869046449661255
7	0.6513226628303528	0.911749541759491	0.5879185795783997	1.0641013383865356
8	0.6511794924736023	0.9085033535957336	0.5872984528541565	1.0686391592025757
9	0.6548218727111816	0.9025095701217651	0.5983810424804688	1.0420494079589844
10	0.6564544439315796	0.8986133337020874	0.5824570059776306	1.0833896398544312
11	0.6578961610794067	0.8947017192840576	0.5879894495010376	1.0990359783172607
12	0.6600415706634521	0.8896796107292175	0.5955328941345215	1.058488130569458
13	0.6607016324996948	0.890235424041748	0.5995482206344604	1.0522098541259766
14	0.6617032885551453	0.8865048885345459	0.5851036310195923	1.084041714668274

Table 2. Training-log in a Simple ML model With Normalization

Table 1은 EEG 데이터를 정규화하지 않고 학습을 했을 때이고, Table 2는 EEG 데이터를 [-1, 1]로 정규화를 하고 학습을 진행한 결과입니다. 학습에 사용한 모델은 챌린지에서 제공한 샘플 모델입니다. 정규화를 통해 약 2-3%의 정확도 향상을 이뤄낼 수 있었습니다. 이후의 학습에서 또한 이 정규화된 데이터를 사용하여 학습을 진행할 예정입니다.

c. An Architecture of Auditory Specialized Encoder

현재 챌린지를 풀기 위한 인코더의 설계는 다음과 같습니다.

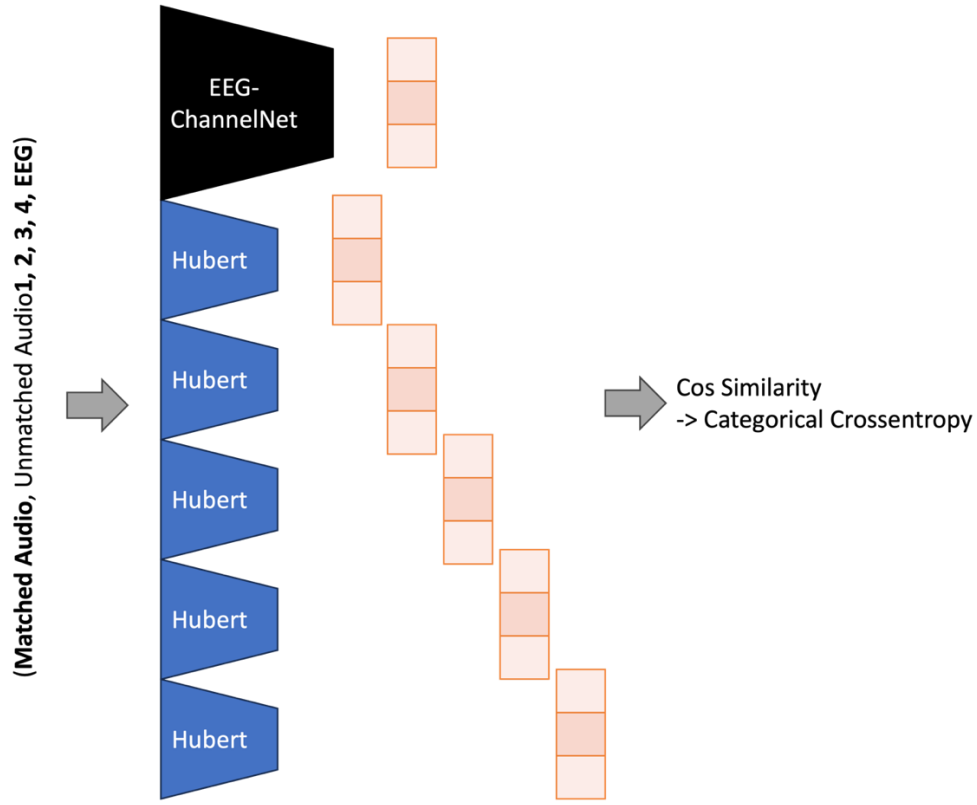

Figure 3. An Architecture of Auditory Specialized Encoder

Input은 EEG 데이터, EEG 데이터와 매치되는 Audio 1개, 불일치 하는 Audio 4개입니다. 현재 본 연구에서 사용하는 Encoder의 종류는 EEG-ChannelNet [9]와 Hubert [16] 입니다. EEG-ChannelNet은 EEG로부터 시각 이미지 재현을 효과적으로 수행하는 것이 입증된 인코더입니다. Hubert는 Self-Supervised 기반의 모델로, 본 설계에서는 Encoder 부분만 가져와 사용하고 있습니다. 각 인코더로부터 Input Data의 Vector를 뽑아내면, 이들의 Cosine Similarity 값을 구하여 유사도를 측정합니다. Loss Function으로는 Categorical Crossentropy를 사용하여 Multi-label 데이터 중 매칭되는 것에 학습되도록 했습니다.

이 챌린지를 통해 EEG-Image 인코더가 EEG-Audio 인코더에도 효과적으로 사용될 수 있음을 입증하고자 합니다. 이후에는 그 반대의 경우도 입증한 후, 두 데이터를 병합해 범용 EEG Encoder를 제작할 계획입니다.

IV. Conclusion

*II.a.*는 현재 가상 설계 단계이고, *II.c.*는 실험 도중에 있는 단계입니다. 아직 많은 epochs를 돌려보진 못했지만, epoch가 증가함에 따라 training accuracy가 90% 이상으로 증가하고 training loss value가 계속적으로 감소함을 확인하였습니다. 이를 토대로 모델이 데이터의 특징은 뽑아내고 있음을 알 수 있습니다. 하지만 현재까지 진행된 epoch가 약 6-7 정도인 점과 Validation Accuracy, Loss Value는 적절한 값을 보여주지 못하고 있는 문제가 발생하고 있습니다. *II.B.*에서 적절한 수정이 가해질 수도 있음을 시사합니다.

본 프로젝트는 궁극적으로 인간의 시각과 청각을 재현할 수 있는 Universal EEG Encoder의

제안을 목표로 합니다. 그 과정에서 (EEG-Image), (EEG-Audio) 데이터의 상관관계를 파악하고자 [3]의 챌린지를 준비하고 있습니다. 챌린지의 결과와 상관없이, 본 연구의 궁극적인 목표에 도움이 될 것이라 기대합니다.

Reference

다음의 목록은 MLA 포맷을 따랐습니다.

[1] Mozafari, Milad, Leila Reddy, and Rufin VanRullen. "Reconstructing natural scenes from fmri patterns using bigbigan." *2020 International joint conference on neural networks (IJCNN)*. IEEE, 2020.

[2] Shen, Guohua, et al. "End-to-end deep image reconstruction from human brain activity." *Frontiers in computational neuroscience* 13 (2019): 21.

[3] Ren, Ziqi, et al. "Reconstructing seen image from brain activity by visually-guided cognitive representation and adversarial learning." *NeuroImage* 228 (2021): 117602.

[4] Khare, Sanchita, et al. "NeuroVision: perceived image regeneration using cProGAN." *Neural Computing and Applications* 34.8 (2022): 5979-5991.

[5] Ye, Zesheng, et al. "Self-supervised cross-modal visual retrieval from brain activities." *Pattern Recognition* 145 (2024): 109915.

[6] Likhosherstov, Valerii, et al. "Polyvit: Co-training vision transformers on images, videos and audio." *arXiv preprint arXiv:2111.12993* (2021).

[7] Gong, Yuan, et al. "Uavm: Towards unifying audio and visual models." *IEEE Signal Processing Letters* 29 (2022): 2437-2441.

[8] "Auditory EEG Challenge 2024." *ICASSP 2024 SP Grand Challenges*, https://exporl.github.io/auditory-eeg-challenge-2024/.

[9] Palazzo, Simone, et al. "Decoding brain representations by multimodal learning of neural activity and visual features." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.11 (2020): 3833-3849.

[10] Karras, Tero, et al. "Progressive growing of gans for improved quality, stability, and variation." *arXiv preprint arXiv:1710.10196* (2017).

[11] Singh, Prajwal, et al. "EEG2IMAGE: Image reconstruction from EEG brain signals." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[12] Accou, Bernd, et al. "SparrKULee: A Speech-evoked Auditory Response Repository of the KU Leuven, containing EEG of 85 participants." *bioRxiv* (2023): 2023-07.

[13] Thornton, Mike, Danilo Mandic, and Tobias Reichenbach. "Relating EEG recordings to speech using envelope tracking and the speech-FFR." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[14] Piao, Zhenyu, et al. "Happyquokka System For Icassp 2023 Auditory Eeg Challenge." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.

[15] Défossez, Alexandre, et al. "Decoding speech perception from non-invasive brain recordings." *Nature Machine Intelligence*(2023): 1-11.

[16] Hsu, Wei-Ning, et al. "Hubert: Self-supervised speech representation learning by masked prediction of hidden units." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 3451-3460.

Appendix

- 졸업프로젝트 회의록.docx