EEGPT: Pretrained Transformer for Universal and Reliable Representation of EEG Signals

상태	완료
Journal	NeurIPS
Year	2024.09
Summary	보편적인 EEG 특징 추출을 위해 설계된 새로운 10M-parameter pretrained transformer 모델인 EEGPT 제안
Limitations	향후 더 큰 데이터셋, 더 긴 시간 길이의 EEG 시그널로 학습 확장 예정
Link	https://proceedings.neurips.cc/paper_files/paper/2024/file/4540d267eeec4e5dbd9dae9448f0b739-Paper-Conference.pdf
category	EEG LLM Pre-training

https://github.com/BINE022/EEGPT

Introduction

- EEG는 비침습적이고 휴대성이 좋아 BCI나 신경과학, 의료에서 널리 사용됨
- 하지만 SNR(Signal-to-Noise Ratio)이 낮고, 사람마다 데이터가 다르며(Inter-subject variability), 특정 작업 의존성이 높기 때문에 일반적인 표현 학습이 어려움
- 기존의 self-supervised learning이 NLP나 CV 분야에서 성과를 보였지만, EEG에서는 효과가 제한적임
  - EEG 표현 학습 프레임워크 : SimCLR 프레임워크를 시계열 데이터에 확장해 채널 특징 추출기 훈련
  - BENDR : masked autoencoder 및 contrastive learning 적용
  - EEG2VEC : contrastive loss 및 reconstruction loss 기반 EEG 표현 학습
  - BIOT : 채널 불일치, 가변 길이 및 결측 값 문제 해결 → 고정 길이 세그먼트로 토큰화
  - LaBraM : 채널 패치 세분화 → neural tokenizer 훈련 → 원래 neural code 예측 위해 transformer 훈련
- 본 논문에서는 **EEGPT라는 10M 파라미터의 트랜스포머 모델**을 제안함:
  - **Spatio-temporal Representation Alignment**와 **Mask-based Reconstruction**이라는 **Dual self-supervised 방식** 사용
  - **계층적 구조**로 공간적, 시간적 정보를 나눠서 처리 → 계산 효율성, BCI 적용 유연성 증가
  - **local spatio-temporal embedding method** 사용 → 다양한 EEG 장치 간 강건성과 호환성 높임
  - 다양한 데이터셋으로 pretrain되어 **범용적인 EEG 표현** 학습 → downstream task에는 **linear-probing method** 사용

Method

- Background
  - 식 1 : masked autoencoder → 식 2 : 학습된 표현 z를 명시적으로 나타내기 위해 spatio-temporal representation alignment 추가 (dual self-supervised method)

$$\min_{\theta,\phi} \mathbb{E}_{x\sim D} H\left(d_\phi(z), x \odot (1-M)\right), \quad z = f_\theta(x \odot M) \tag{1}$$

$$\min_{\theta,\phi} \mathbb{E}_{x\sim D} H\left(d_\phi(z), x \odot (1-M)\right) + H(z, f_\theta(x)), \quad z = f_\theta(x \odot M) \tag{2}$$

▼ 📚 **식 설명**

- ⊙ : element-wise product
- M : 패치 마스크
- $f_\theta(\cdot)$와 $d_\phi(\cdot)$: 각각 인코더와 디코더
- z : 학습된 표현
- $H(\cdot,\cdot)$: 유사성 측정 방법
- 손실 함수를 최소화함으로써 모델은 입력 신호의 최적 표현 z를 학습함 → but, BERT 계열 모델에는 명시적인 표현 z가 없음 → why? 인코더 디코더 분할 없기 때문
- 식 2로 바꿈으로써, 인코딩된 표현이 다소 더 넓은 의미를 갖도록 함

- EEG Pretrained Transformer (EEGPT)
  - 입력 EEG 신호 $x \in \mathbb{R}^{M\times T}$ (M 채널 및 T 시간 포인트)를 **패치 $p_{i,j}$로 나눔**
  - 각 패치를 local spatio-temporal embedding을 통해 토큰 $token_{i,j}$**으로 임베딩** (마스크된 부분 M과 비마스크 부분 $\overline{\mathcal{M}}$로 분리하는 작업 포함)
  - spatio-temporal representation alignment와 mask-based reconstruction을 포함한 dual self-supervised learning을 사용하여 **모델 사전 훈련**
  - **다운스트림 작업**에 linear-probing method 사용

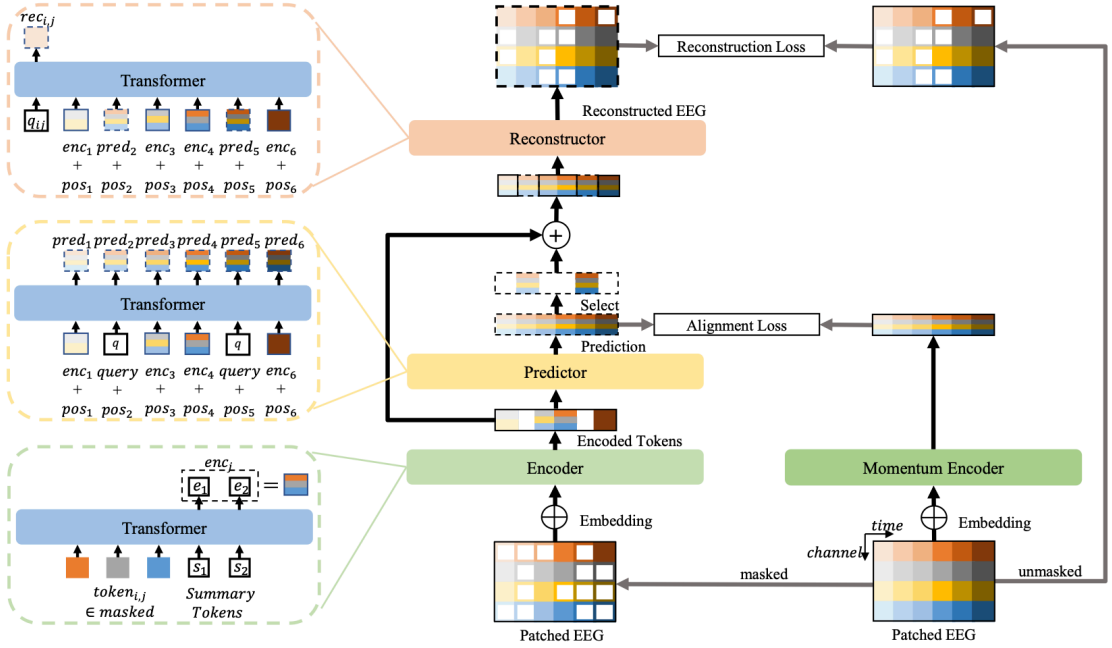

Figure 1: The EEGPT structure involves patching the input EEG signal as $p_{i,j}$ through masking (50% time and 80% channel patches), creating masked part $\mathcal{M}$ and unmasked part $\overline{\mathcal{M}}$ and then embedding as $token_{i,j}$ by local spatio-temporal embedding. The encoder processes the masked part, extracting features (enc_j) consisting of $\{e_i\}_{i=1}^{S}$ for each time segment in the $\mathcal{M}$ part with summary tokens $\{s_i\}_{i=1}^{S}$. The predictor predicts features ($pred_j$) for all time segments, aligning with the Momentum Encoder output ($menc_j$). Based on features extracted by the predictor and encoder, the reconstructor generates $rec_{i,j}$ to reconstruct the EEG signal of the $\overline{\mathcal{M}}$ part.

Spatio-temporal Representation Alignment

- 일반적인 BERT-style 모델은 마스킹된 데이터를 복원하면서 학습함
- EEGPT는 여기서 한 걸음 더 나아가 **복원만 하지 말고, 전체 시그널과 잘 align되도록** 학습
- 이를 위해 **Momentum Encoder** 구조를 도입하여, 안정적인 feature alignment를 유도함
- Encoder : 마스크된 패치에서 공간 정보 통합
- Predictor
  - 인코더로부터 마스킹된 부분의 특징(enc_j)과 시간 위치 정보(pos_j)를 결합하여 전체 인코딩된 특징 예측
  - $\overline{\mathcal{M}}$에 속하는 예측 특징을 생성하기 위해 학습 가능한 벡터 쿼리가 쿼리 토큰으로 사용
- Momentum Encoder
  - $token_{i,j}$를 시간 j의 입력으로 처리하고, 해당 출력 $menc_j$를 생성
  - Mean Square Error(MSE) 기반 alignment loss 적용 → 시공간 표현 정렬 달성 !

Mask-based Reconstruction

- Reconstructor가 생성한 **재구성 패치를 M 부분의 원시 패치 $p_{i,j}$와 정렬**
- Reconstructor
  - 인코더에 의해 인코딩된 M 부분의 특징 enc_j와 예측기에 의해 예측된 M 부분의 특징 pred_j와 함께 시간 위치 pos_j를 활용하여 재구성된 패치 $\mathrm{rec}_{u,t}$ 생성
  - 인코더와 reconstructor 사이에 skip connection 설정 : 피처 유지 및 수렴 속도 높이는데 도움
- Reconstruction Loss (Mean Squared Error)를 사용함

Local Spatio-Temporal Embedding

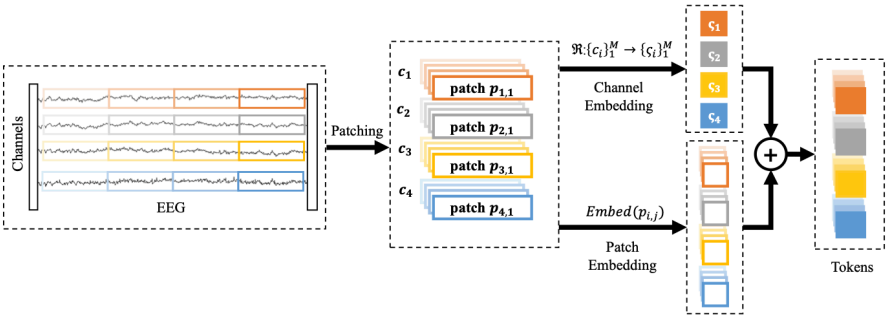

Figure 2: Illustration of local spatio-temporal embedding. The EEG signal is divided into equally sized patches in the spatio-temporal dimensions. Each patch represents a time segment for a specific channel without overlap. The patches are linearly embedded and incorporated with channel embedding information to obtain a corresponding feature.

- EEG 신호를 시공간 차원에서 동일한 크기의 패치 $p_{i,j}$로 나누고,
- 채널 임베딩 정보와 결합하여 선형적으로 임베딩됨
  - **채널 임베딩 벡터와 codex book 매핑** → 채널을 모델 입력의 채널에 유연하게 대응할 수 있음
- 이렇게 하면 **다양한 장비와 채널 수**에도 잘 적응할 수 있는 표현을 학습함

Linear-Probing Method

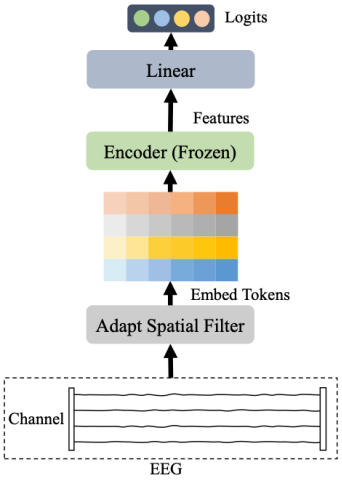

Figure 3: Linear-probing method.

- Downstream task에서는 사전 학습된 인코더는 freeze하고, 그 위에 간단한 Linear Classifier만 학습함
  - 추가 선형 모듈의 매개변수만 변경하는 linear-probing method 도입
- 이렇게 하면 과적합을 방지하고, **Encoder의 표현력만 평가 가능**함

Experiments

Datasets and Data Processing

Table 1: Datasets for pretraining and downstream tasks

	Datasets	Paradigms	Subjects	Targets
pretraining Datasets	PhysioMI	MI&ME	109	5
	HGD	MI	14	4
	TSU	SSVEP	35	40
	SEED	EMO	15	3
	M3CV	MULTI	106	-
Downstream Datasets	BCIC-2A	MI	10	4
	BCIC-2B	MI	10	2
	Sleep-EDFx	SLEEP	197	5
	KaggleERN	ERN	26	2
	PhysioP300	P300	9	2
	TUAB	Abnormal	2383	2
	TUEV	Event	288	6

- 다양한 paradigm의 EEG 데이터셋을 사용:
  - Pretraining: PhysioMI, HGD, TSU, SEED, M3CV
  - Downstream: BCIC-2A/2B, Sleep-EDFx, KaggleERN, TUAB, TUEV 등
- 공통적인 전처리 수행: 밴드패스 필터링, 리샘플링, 정규화 등

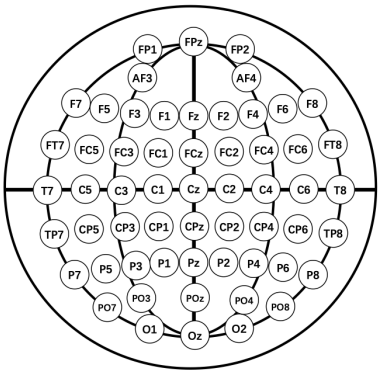

Figure 4: Electrode locations.

Implementation and Settings

- encoder, predictor, reconstructor는 **Vision Transformer(ViT)** 구조 기반
- 훈련하는 동안 50% time, 80% channel 마스킹 사용
- AdamW + OneCycle 학습률 스케줄러 사용

Downstream Experiment Results

- EEGPT는 다양한 다운스트림 태스크에서 기존 모델(BENDR, BIOT, LaBraM)보다 **더 높은 성능**을 달성
- EEG 데이터의 문제를 해결한다는 것을 보여줌

Table 2: The results of different methods on TUAB.

Methods	Model Size	Balanced Accuracy	AUROC
SPaRCNet [44]	0.79M	0.7896±0.0018	0.8676±0.0012
ContraWR [45]	1.6M	0.7746±0.0041	0.8456±0.0074
CNN-T [46]	3.2M	0.7777±0.0022	0.8461±0.0013
FFCL [47]	2.4M	0.7848±0.0038	0.8569±0.0051
ST-T [48]	3.5M	0.7966±0.0023	0.8707±0.0019
BIOT [15]	3.2M	0.7959±0.0057	0.8815±0.0043
Ours-Tiny	4.7M	0.7959±0.0021	0.8716±0.0041
Ours	25M	0.7983±0.0030	0.8718±0.0050

Table 3: The results of different methods on TUEV.

Methods	Model Size	Balanced Accuracy	Weighted F1	Cohen's Kappa
SPaRCNet [44]	0.79M	0.4161±0.0262	0.7024±0.0104	0.4233±0.0181
ContraWR [45]	1.6M	0.4384±0.0349	0.6893±0.0136	0.3912±0.0237
CNN-T [46]	3.2M	0.4087±0.0161	0.6854±0.0293	0.3815±0.0134
FFCL [47]	2.4M	0.3979±0.0104	0.6783±0.0120	0.3732±0.0188
ST-T [48]	3.5M	0.3984±0.0228	0.6823±0.0190	0.3765±0.0306
BIOT [15]	3.2M	0.5281±0.0225	0.7492±0.0082	0.5273±0.0249
Ours-Tiny	4.7M	0.5670±0.0066	0.7535±0.0097	0.5085±0.0173
Ours	25M	0.6232±0.0114	0.8187±0.0063	0.6351±0.0134

Table 4: The results of universal EEG models on various datasets.

Datasets	Methods	Balanced Accuracy	Cohen's Kappa	Weighted F1 / AUROC
BCIC-2A	BENDR	0.4899±0.0070	0.3199±0.0094	0.4836±0.0076
	BIOT	0.4590±0.0196	0.2787±0.0261	0.4282±0.0289
	LaBraM	0.5613±0.0052	0.4151±0.0069	0.5520±0.0052
	Ours	**0.5846±0.0070**	**0.4462±0.0094**	**0.5715±0.0051**
BCIC-2B	BENDR	0.7067±0.0011	0.4131±0.0022	0.7854±0.0029
	BIOT	0.6409±0.0118	0.2817±0.0236	0.7095±0.0141
	LaBraM	0.6851±0.0063	0.3703±0.0125	0.7576±0.0067
	Ours	**0.7212±0.0019**	**0.4426±0.0037**	**0.8059±0.0032**
Sleep-EDFx	BENDR	0.6655±0.0043	0.6659±0.0043	0.7507±0.0029
	BIOT	0.6622±0.0013	0.6461±0.0017	0.7415±0.0010
	LaBraM	0.6771±0.0022	0.6710±0.0006	0.7592±0.0005
	Ours	**0.6917±0.0069**	**0.6857±0.0019**	**0.7654±0.0023**
KaggleERN	BENDR	0.5672±0.0020	0.1461±0.0037	0.6030±0.0044
	BIOT	0.5118±0.0089	0.0297±0.0224	0.5495±0.0167
	LaBraM	0.5439±0.0029	0.0944±0.0066	0.5693±0.0052
	Ours	**0.5837±0.0064**	**0.1882±0.0110**	**0.6621±0.0096**
PhysioP300	BENDR	0.6114±0.0118	0.2227±0.0237	0.6588±0.0163
	BIOT	0.5485±0.0325	0.0968±0.0647	0.5308±0.0333
	LaBraM	0.6477±0.0110	0.2935±0.0227	0.7068±0.0134
	Ours	**0.6502±0.0063**	**0.2999±0.0139**	**0.7168±0.0051**

Ablation Experiment Results

Table 5: The results of the ablation study.

Variants	$\mathcal{L}_A$	$\mathcal{L}_R$	BCIC-2A-BAC	BCIC-2B-AUROC	KaggleERN-AUROC
A: w/o $\mathcal{L}_A$	37.13	0.57	0.5287±0.0086	0.7264±0.0381	0.5752±0.0164
B: w/o LN	0.15	0.002	0.5567±0.0088	0.7920±0.0012	0.5891±0.0227
C: w/o skip	0.12	0.56	0.5796±0.0011	0.7702±0.0122	0.6356±0.0296
D: with all	0.24	0.56	**0.5846±0.0070**	**0.8059±0.0032**	**0.6621±0.0096**

- Spatio-temporal alignment, LayerNorm, skip connection 등을 하나씩 제거해본 결과
  - 각각의 구성요소가 성능 향상에 **실질적인 기여**를 하고 있음이 확인됨

Pretrain Experiment Results

- 모델의 크기(파라미터 수)가 증가할수록 성능도 상승 (Scaling law 확인)
- Summary token 수, embedding 차원 등도 성능에 영향을 줌

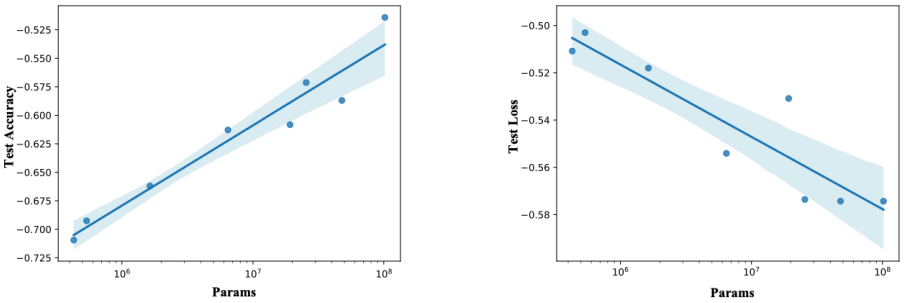

Figure 5: Scaling laws with EEGPT parameter size N. Axes are all on a logarithmic scale.

Conclusion

- EEGPT는 spatio-temporal alignment + reconstruction이라는 **dual self-supervised 방식**으로 EEG의 robust한 표현을 학습함
- 다양한 데이터셋과 task에 대해 **범용적이고 신뢰할 수 있는 성능**을 보임
- 향후 더 큰 데이터셋, 더 긴 시간 길이의 EEG 시그널로 학습 확장 예정