Exploring Large-Scale Language Models to Evaluate EEG-Based Multimodal Data for Mental Health

상태	완료
Journal	UbiComp
Year	2024.10
Summary	mental health 분야에서 multimodal 데이터를 활용한 MultiEEG-GPT 제안 및 평가
Limitations	fine-Tuning이나 multi-strategy hierarchical 등을 통해 MultiEEG-GPT 개선 가능(+ responsible AI)
Link	https://arxiv.org/pdf/2408.07313
category	EEG LLM Multimodal Prompt-tuning

mental health 분야에서 multimodal 데이터를 활용한 MultiEEG-GPT 제안 및 평가 : prompt engineering, multimodal data(EEG, facial expression, audio)

키워드

- Mental Health
- EEG
- Large Language Model
- Prompt Engineering

Introduction

health condition을 정확하게 측정하고 분류하려면 psychological evaluation이 필요함. 일반적으로 EEG, HRV, EDA(Electrodermal Activity, 피부전도도)와 같은 생리적 신호는 신뢰성과 정확한 식별성으로 인해 mental health assessments로 많이 사용됨.

LLM은 multimodal 데이터 처리, human-in-the-loop와 같은 상호 작용적 커뮤니케이션 방법을 통한 health agents 생성, 그리고 비용 절감을 위한 일반 모델 기반의 도메인 지식 fine-tuning을 포함한 여러 가지 장점을 제공함. But, mental health LLM 분야에서는 단일 모달만을 사용해 특정 작업에만 초점을 맞춤. multimodal LLM도 존재하지만 audio와 video 정도만 사용함. mental health assessment에서는 EEG 및 다른 생리적 신호가 중요한 역할을 하기 때문에 해당 데이터를 반영한 LLM 필요함.

따라서, EEG, facial expression, audio(text)와 같은 multimodal을 사용해 mental health를 평가하는 방법인 MultiEEG-GPT를 제안함. GPT-4o 기반이며 다양한 정신 건강 상태를 분류하는 데 있어 multimodal LLM의 능력을 이해하는 연구임.

- contributions
  - 다양한 건강 상태를 인식하는데 있어 multimodal을 사용한 MultiEEG-GPT의 예측 능력 평가를 위해 zero- / few-shot 프롬프트 설계
  - MultiEEG-GPT의 유효성을 검증하기 위해 3가지 데이터베이스에서 실험
  - multimodalities가 건강 상태 예측을 어떻게 향상시키는지 이해하기 위한 심층 분석

future work으로 social robot과 통합 개발하는 방법 제안함. (호오? 고냥 페퍼에 이거 연결해서 사용자 테스트 하고 분석해봐도 좋을 듯)

Related Work

머신러닝(k-NN,m SVM) EEG 분류 정확도 67.07%, 딥러닝(CNN, BiLSTM) EEG 분류 정확도 93.20% 정도임.

Methodology

Dataset Selection

- MODMA [link] ; 우울장애 환자와 대조군의 구술 기록(오디오) 및 EEG(이미지로 변환 가능)로 구성됨. 우울 장애 진단에 대한 이진 라벨.
- PME4 [link] ; 오디오, 비디오(공개되지 않음), EEG, EMG(electromyography)에 대한 멀티모달 감정 데이터셋. 11명의 연기 전공 학생으로부터 수집되었고, 7가지 감정이 라벨임.
- LUMED-2 [link] ; 다양한 자극에서 중립, 행복, 슬픔을 인식하고 분류하기 위해 facial expression, EEG, GSR(galvanic skin response, 피부 전도 반응)로 구성됨.

Prompt Design

Table 1: The zero-shot and few-shot prompting strategies. <MOD1>, <MOD2> and <MOD3> as placeholders denote three different modalities. XXX is the description of collection and visualization process. <SYM> as a placeholder denotes the symptom to be diagnosed. For example, for depression analysis <SYM> should be replaced with depression. The example is for mental health diagonsis with three classes. The label description "0 denotes XXX" of the classes could be added or removed to accommodate for more or less classes.

Role-play prompt	Imagine you are a mental health expert expert at analyzing the emotion and mental health status.
Task specification	The below is <MOD1>, <MOD2> and <MOD3> data. <MOD1> data is collected through XXX and visualized in XXX form. <MOD2> data is collected through XXX and visualized in XXX form. <MOD3> data is collected through XXX and visualized in XXX form. Analyze the <SYM> status of the person. 0 denotes XXX, 1 denotes XXX, 2 denotes XXX.
Rules	[Rule]: Do not output other text.

few-shot은 위 프롬프트에서 샘플 한개 추가함. 대신, task specification에서 다양한 후보 클래스 레이블을 제공하는 대신 올바른 클래스 레이블을 제공함.

Experiment

Settings

Dataset Settings. EEG 신호 처리를 위해 MNE 라이브러리를 사용함. 대역 통과 필터는 firwin window design 사용하여 적용함. 필터링된 데이터는 평균참조로 재참조함. 데이터셋 길이는 530s, 5s 및 1.65-4.15s로 무작위로 설정된 유도 시간을 고려하고 10개의 균등 간격으로 샘플링함. 얼굴 표정은 비디오의 중간 프레임(ex. 비디오 길이가 10초인 경우, 정확히 5초 타임스탬프에서 프레임 선택) 또는 이미지에서 선택함. 오디오는 GPT-4o에서 오디오 입력을 지원하지 않았기 때문에, audio feature와 text를 입력으로 사용함. librosa 라이브러리를 사용해 MFCC, 멜 스펙트로그램, 크로마 STFT 등의 feature 추출하고 text 형식으로 나타냄. text의 경우 가장 크고 최신 모델이며 의료 작업에서 사용된적 있는 ASR 시스템 사용하여 기록함.

Model Settings. 모든 태스크는 다중 클래스 분류 문제임. GPT-4o로 진행했고, 매 반복마다, 해당 데이터셋에서 무작위로 샘플 하나를 선택하여 1-shot 샘플로 사용함. 모든 zero- / few-shot 실험에 대해, 각 데이터셋에서 5회 테스트하고 평균 정확도와 표준 편차를 봄.

Results and Discussions

Multimodal analysis. Figure 1은 LUMED-2 및 PME4 데이터셋을 사용한 2개의 zero-shot 케이스 예제임. (a)에서 MultiEEG-GPT는 먼저 이미지를 처리하고 이후 EEG 토폴로지 맵을 분석함. 결과는 중립적인 기분이라고 출력했고 정답임. (b)에서 MultiEEG-GPT는 먼저 참가자의 오디오 특징을 분석한 후, 토폴로지 맵을 통해 EEG 특징을 분석함. 마지막으로 다양한 특징을 결합하여 참가자가 슬픈 기분임을 예측했고 정답임.

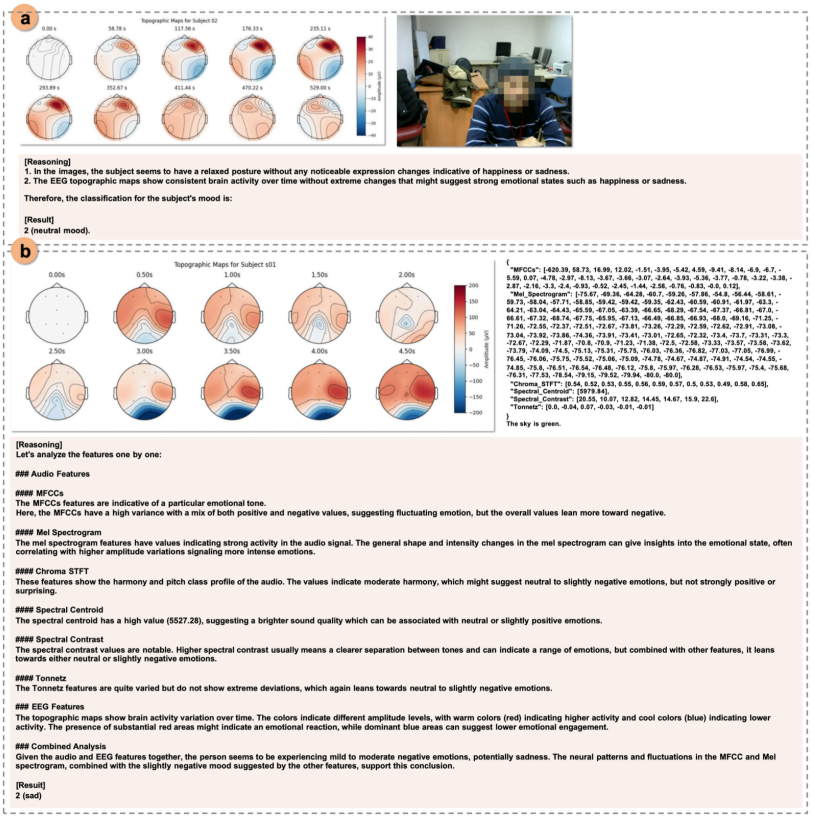

Figure 1: Case analysis for LUMED-2 and PME4 datasets (the person's face has been blurred for ethical reasons). Figure (a) illustrates one subject's input EEG topology map and his facial expression, as well as the prediction result and the text explanation from LUMED-2 dataset. Figure (b) illustrates one subject's input EEG topology map, audio features, input audio transcription "The sky is green.", as well as the prediction result and the explanation, from PME4 dataset. In both cases, the model makes the accurate predictions when processing both modalities.

이는 MultiEEG-GPT가 (1) 각 modality를 별도로 분석하고 (2) 다양한 modality의 출력을 공동으로 집계할 수 있는 능력이 있음을 보여줌. 또한, 단일 modality로는 기분을 정확히 식별할 수 없다는 것이 분명함. 예를 들어, MultiEEG-GPT는 (b)의 상태를 "an emotional reaction"으로 식별했으나 EEG 특징으로 참가자가 슬프다고 정확히 명시하지는 않았음. 오디오 특징과 EEG 특징을 결합하여 MultiEEG-GPT는 정확한 예측을 달성함.

Performance of MultiEEG-GPT. 제안된 모델은 세 데이터베이스에서 각각 최고의 단일 modality 성능에 비해 multimodal의 성능이 더 좋음을 보여줌. 이는 오디오 및 비디오와 같은 일반적으로 사용되는 모달리티에 추가하여 EEG 데이터를 포함하는 것의 중요성을 강조함.

Table 2: Ablation experiment on 3 different multimodal data (EEG image, facial expression, audio). The line with no EEG image, facial expression, audio was determined through majority voting. For few-shot prompting, we chose M=1, which meant we added one few-shot sample in the prompt.

Prediction Accuracy (%)						
Strategy	EEG	Facial Expression	Audio	MODMA	PME4	LUMED-2
Zero-shot Prompting	×	×	×	$50.0_{\pm 0.00}$	$14.28_{\pm 0.00}$	$33.33_{\pm 0.00}$
	✓	×	×	$53.79_{\pm 2.46}$	$21.05_{\pm 1.71}$	$34.61_{\pm 1.28}$
	×	✓	×	–	–	$38.46_{\pm 1.54}$
	×	×	✓	$69.35_{\pm 2.53}$	$15.38_{\pm 1.42}$	–
	✓	✓	×	–	–	**$46.13_{\pm 2.42}$**
	✓	×	✓	**$73.54_{\pm 2.03}$**	**$28.57_{\pm 2.41}$**	–
Few-shot Prompting (M=1)	×	×	×	$50.0_{\pm 0.00}$	$14.28_{\pm 0.00}$	$33.33_{\pm 0.00}$
	✓	×	×	$62.71_{\pm 3.23}$	$26.00_{\pm 1.78}$	$36.37_{\pm 1.62}$
	×	✓	×	–	–	$43.64_{\pm 1.85}$
	×	×	✓	$69.92_{\pm 1.53}$	$19.13_{\pm 1.29}$	–
	✓	✓	×	–	–	**$52.73_{\pm 2.16}$**
	✓	×	✓	**$79.00_{\pm 1.59}$**	**$37.00_{\pm 2.30}$**	–

zero- / few-shot 프롬프트 적용했을 때, few-shot이 zero보다 높은 성능을 보임. 이는 추가 샘플이 인식을 향상시킨다는 것을 시사하며, 이전 연구 결과와 일치함. 추가 샘플은 feature 비교의 기준 역할을 할 가능성이 있음. 결과는 추가 샘플의 일반적인 이점을 나타내며, one-shot 프롬프트 설정에서 특정 샘플이 의도적으로 선택되지 않았음을 강조함.

(개인적으로 facial + audio 조합 없는게 아쉬움.. 2개 조합으로도 성능이 충분히 좋으면 굳이 EEG를 쓰는 이유가 없음.. 아마 저자들이 실험은 했는데,, 성능이 EEG 포함한것보다 더 좋아서 뺀 걸수도?)

Conclusion and Future Work

향후 mental health를 위한 multimodal LLM 활용에 대한 Instruction Fine-Tuning이나 multi-strategy hierarchical prediction과 같은 전략을 통해 개선할 수 있는 상당한 가능성이 있음. + responsible AI 강조