TRIBE: TRImodal Brain Encoder for whole-brain fMRI response prediction

https://www.arxiv.org/pdf/2507.22229

0. Introduction

- 실제 인간 뇌는 시각, 언어, 청각 등 다중 감각 정보를 동시에 처리함.
- 그러나 기존 인코딩 모델은 이러한 멀티모달 자극 통합 표현을 충분히 반영하지 못함.
- 텍스트, 이미지, 오디오를 통합하는 Trimodal Brain Encoder (RIBE) 제안
- 멀티모달 자극 기반으로 전뇌 fMRI 반응을 동시에 예측하는 구조 설계
- 기존 단일·이중 모달 인코딩 모델 대비 예측 정확도 향상

1. Overview

- 서로 다른 자극 모달리티에서 추출한 표현을 공통 잠재 공간에서 융합하여 전뇌 voxel 단위 fMRI 반응을 직접 회귀 예측
- 기본 구조
  - 이미지 인코더 + 텍스트 인코더 + 오디오 인코더로 구성된 3개 모달리티 인코딩
  - 모달리티별 feature를 통합하는 cross-modal fusion 모듈
  - 최종 Brain Encoder가 전체 뇌 voxel 반응 벡터 출력
- 적용 범위

  자연 영상, 설명 텍스트, 음성 자극을 포함한 fMRI 데이터셋에서 전뇌 반응 예측 수행

2. Challenges

- 멀티모달 정렬 문제

서로 다른 모달리티 feature를 동일한 뇌 반응 공간에 정렬하는 것이 어려움

- 전뇌 스케일 문제

  voxel 수가 매우 많아 고차원 회귀 문제로 인한 학습 불안정 및 과적합 위험 존재

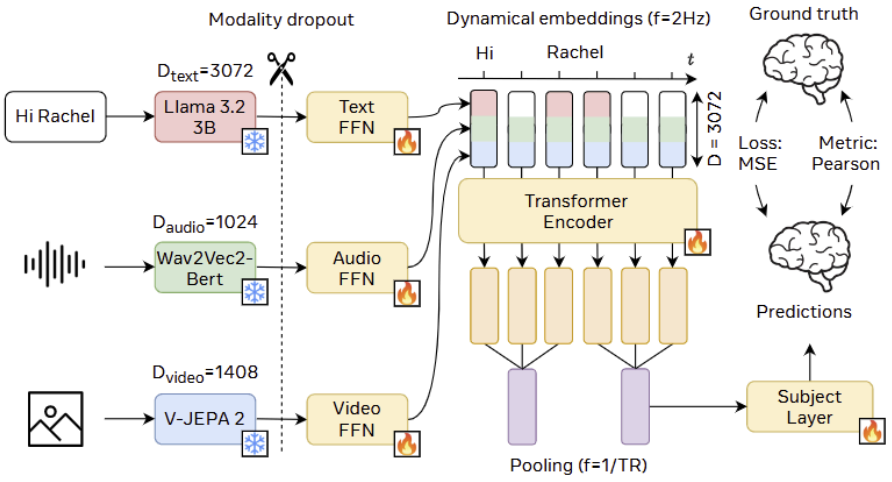
- 자극-뇌 반응 비선형성

  자극 의미와 뇌 반응 사이의 관계가 단순 선형 모델로 설명되기 어려움

- 데이터 한계

  fMRI 데이터 규모 제한 및 노이즈 문제로 일반화 성능 확보가 어려움

3. Method



- Trimodal Input Encoding
  - Image Encoder
    - 사전학습된 비전 트랜스포머 또는 CNN 기반 특징 추출
  - Text Encoder
    - 사전학습 언어 모델 기반 문장·설명 임베딩 생성
  - Audio Encoder
    - 음성 또는 환경음 특징 추출용 오디오 인코더 사용
- Cross-Modal Fusion
  - 세 모달리티 임베딩을 공통 잠재 공간으로 투영

    - attention 또는 gating 기반 융합 모듈로 통합 표현 생성
- Brain Encoder
    - 통합된 멀티모달 표현을 입력으로 받아
    - 전뇌 voxel 단위 fMRI 반응을 동시에 예측하는 회귀 네트워크 구성
- Training Strategy
    - 자극 입력과 실제 fMRI 반응 간 MSE 또는 상관 기반 손실 함수 사용
    - subject별 또는 subject-agnostic 설정으로 학습

4. Experiments

- 사용 데이터

  자연 자극 기반 fMRI 데이터셋 사용 (Natural Scenes Dataset 가능성 높음, 추측입니다)
- 실험 설정
    - 비교 대상:
        - 단일 모달 인코딩 모델
        - 기존 dual-modal brain encoding 모델
    - 평가 지표:
        - voxel-wise Pearson correlation
        - prediction accuracy 또는 explained variance
- 추가 실험
    - 모달리티 제거 실험으로 각 입력 기여도 분석
    - 뇌 영역별 성능 비교 (시각 피질, 언어 영역 등)

5. Results

Rank	Team	Mean score	Subject 1	Subject 2	Subject 3	Subject 5
1	**Ours**	**0.2146**	0.2381	0.2105	0.2377	0.1720
2	NCG	**0.2096**	0.2353	0.2046	0.2268	0.1718
3	SDA	**0.2094**	0.2233	0.2072	0.2271	0.1798
4	MedARC	**0.2085**	0.2295	0.2003	0.2300	0.1743
5	CVIU-UARK	**0.2055**	0.2306	0.2010	0.2240	0.1662

Table 1: **Our model achieves first place in the Algonauts 2025 public leaderboard.** We report the results of the top five out of 263 teams.

OOD	Movie	Mean score	Subject 1	Subject 2	Subject 3	Subject 5
✗	Friends Season 7	**0.3195**	0.3419	0.3239	0.3346	0.2775
✓	Pulp Fiction	**0.2604**	0.2765	0.2611	0.2431	0.2610
✓	Princess Mononoke	**0.2449**	0.2816	0.2507	0.2851	0.1623
✓	Passe-partout	**0.2323**	0.2763	0.2587	0.2370	0.1573
✓	World of Tomorrow	**0.1924**	0.2210	0.1606	0.2196	0.1686
✓	Planet Earth	**0.1886**	0.1483	0.2029	0.2331	0.1699
✓	Charlie Chaplin	**0.1686**	0.2249	0.1289	0.2080	0.1128

Table 2: **Our model generalizes to highly out-of-distribution movies.** We provide the detailed results on the held-out datasets of the Algonauts 2025 competition.

- 주요 성능 결과
  - 기존 단일/이중 모달 모델 대비 전뇌 평균 상관계수 유의미하게 향상
  - 멀티모달 입력을 모두 사용할 때 최고 성능 달성
- 영역별 분석
  - 시각 피질에서는 이미지 모달리티 기여도가 큼
  - 언어·연합 영역에서는 텍스트 모달리티 기여도 증가
  - 청각 관련 영역에서 오디오 모달리티 효과 확인
- Ablation 결과
  - 단일 모달 제거 시 성능 하락
  - 세 모달 통합 시 가장 안정적인 전뇌 예측 성능 확보

6. Insight

- 멀티모달 자극 통합은 전뇌 fMRI 인코딩 성능을 실질적으로 향상시킴
- 기존 시각 중심 뇌 인코딩 모델을 범용 멀티모달 뇌 모델로 확장 가능성 제시
- 인간 인지 처리의 다중 감각 통합 특성을 모델 구조에 반영했다는 점에서 의미 있음
- 모델 구조가 복잡하여 학습 비용과 메모리 사용량 큼

- subject 간 일반화 성능은 제한적일 가능성 존재
- 더 많은 모달리티 (영상, 행동 로그 등) 통합