Прогнозирование снимков FMRI по просмотренному видео

Создание интеллектуальных систем, МФТИ

Барабанщикова Полина, Протасов Дмитрий, Шульган Никита

20 декабря 2022

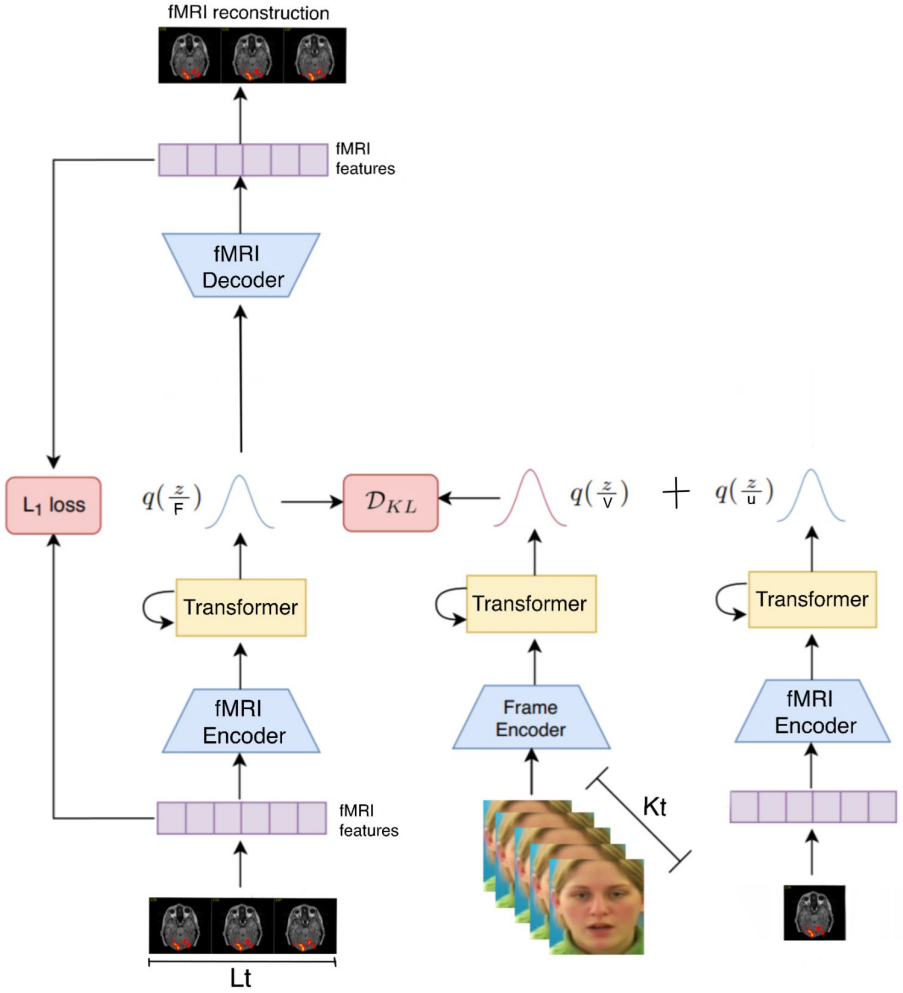
Постановка задачи

Пусть V – видеопоток размерности Kt x C x H x W, где K – число кадров в t секунд, а C, H, W – количество каналов, высота и ширина изображения. И пусть F – fMRI сигнал размерности Lt x X x Y x Z, где L – число измерений за t секунд, а (X, Y, Z) – размерность одного измерения. Также для каждой пары (V, F) известна метка испытуемого u, которая является одним кадром fMRI.

Задача состоит в предсказании fMRI сигнала F по паре (V, u).

Обзор решения

- Вариационный автоэнкодер для предсказания fMRI
- Априорное вариационное распределение генерируется на основе видеопотока и метки пользователя
- VAE и энкодер для видео оснащены механизмом attention для учета временных зависимостей



Функция потерь

$$\mathcal{L}(\theta, \psi) = \text{RecLoss}(H_\theta(F), F) + D_{KL}[q_{\psi_f}(z|F)||q_{\psi_v}(z|V) + q_{\psi_u}(z|u)]$$

- Аналог ELBO с априорным распределением, построенным по видео и метке пользователя
- Первая часть ELBO, отвечающая за реконструкцию сигнала, заменяется на комбинацию L1 лоссов

Детали реализации: данные

Обзор статьи

Некоторые характеристики датасета

- 30 участников с данными FMRI
- 13 чередующихся эпизодов по 30 секунд (речь / музыка)
- Всё переведено на голландский
- Аудио и видеодорожка аннотированы (время начала и конца появления каждого объекта)
- Снимки FMRI в формате BIDS
- Папка видео → 135 tsv файлов (для 129 объектов, и 6 персонажей)

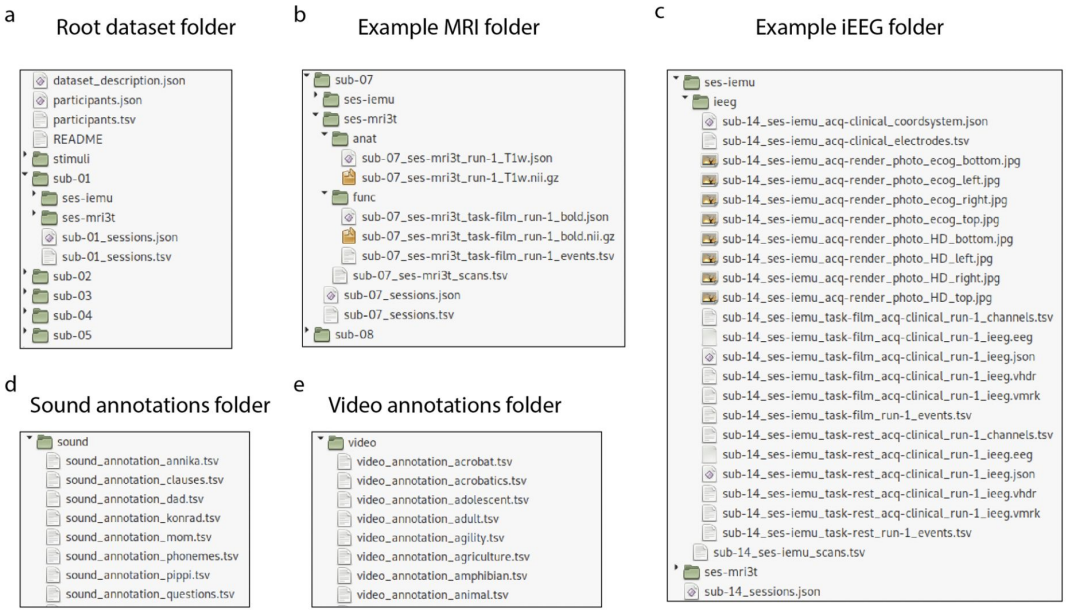


Fig. 2 Overview of data records. (**a**–**e**) Structure of folders and files in the dataset with example (f)MRI and iEEG folders from one subject.

Датасет из статьи

1. Данные
   - 389.728 секунд видео = 9750 фреймов (24 кадра в секунду, размер кадра 480x640x3)
   - 641 измерение fMRI для пользователя (примерно 1.644 измерений в секунду, 40x64x64)
2. Соотношения
   - 3.04 секунды соответствуют 5 измерениям fMRI и 76 фреймам видео
   - Интервал fMRI вида 5*i:5*(i+k) соответствует интервалу видео 76*i:76*(i+k), где i >= 0, k > 0 – целые числа

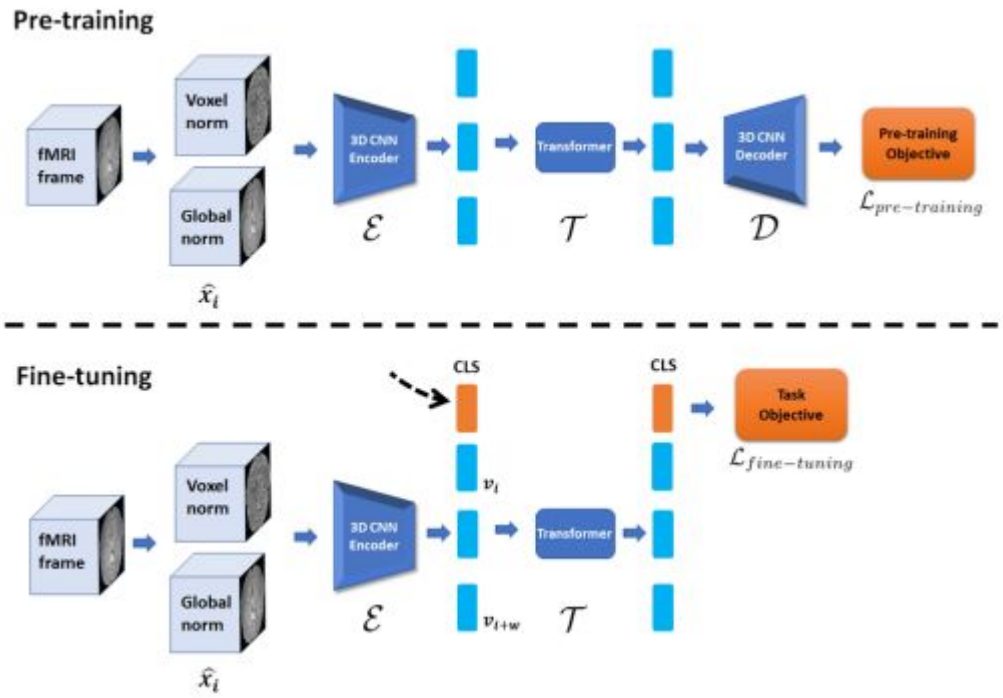
Датасеты для обучения

1. Предобработка
   - по каждому снимку строятся voxel norm и global norm, которые сохраняются в отдельный файл
   - кадры видео обрезаются до размера 224x224
2. Датасет с fMRI
   - элемент датасета – 2 нормализованные версии для T измерений fMRI
   - каждый элемент является тензором размером (2, 40, 64, 64, T)
   - элементы подгружаются из памяти динамически
3. Датасет с fMRI и видео
   - элемент датасета – 2 нормализованные версии для T(+1) измерений fMRI, N соответствующих фреймов видео, индексы позиций для видео
   - пользователи поделены на трейн и тест
   - берется каждый i-ый кадр видео

Детали реализации: модели

TFF. Self-supervised transformers for FMRI representation

1. pretraining on reconstruction 3D-FMRI data (2 phases)
2. fine-tune on specific tasks and show SOTA performance
3. Architecture:
   - preprocessing (voxel, global)
   - 3D-CNN encoder
   - 2 layer transformer
   - 3D-CNN decoder
4. Loss = L1-rec-loss + intensity-loss + perceptual-loss



Video Transformer Network

(https://arxiv.org/pdf/2102.00719.pdf)

Обзор статьи

- Решается задача классификации видео
- Spatial backbone – любая сеть для обработки 2D изображений, используется для извлечения признаков (в оригинальной имплементации: Vision Transformer)
- Temporal attention-based encoder – модель трансформера с механизмом attention, учитывающая временные зависимости (в оригинальной имплементации: Longformer)
- MLP-Head – полносвязная сеть для классификации

Детали реализации

- VTN принимает на вход кадры размера 224x224
- Код и предобученные на Kinetics веса доступны в [https://github.com/bomri/SlowFast/blob/master/projects/vtn/README.md.](https://github.com/bomri/SlowFast/blob/master/projects/vtn/README.md)

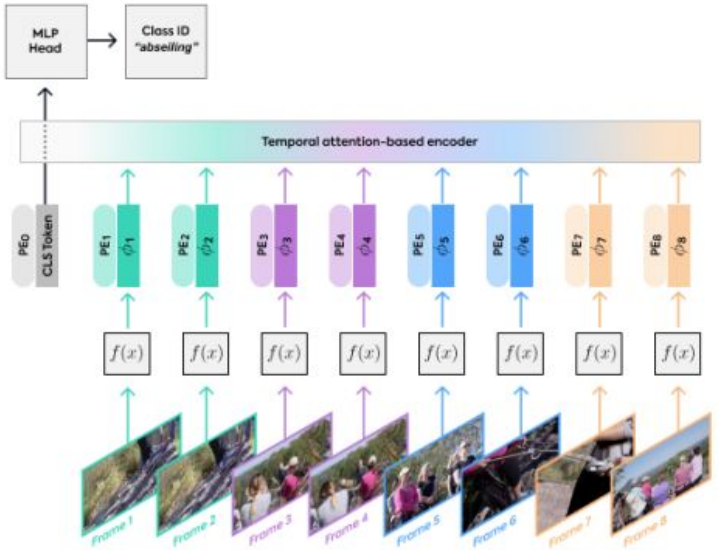
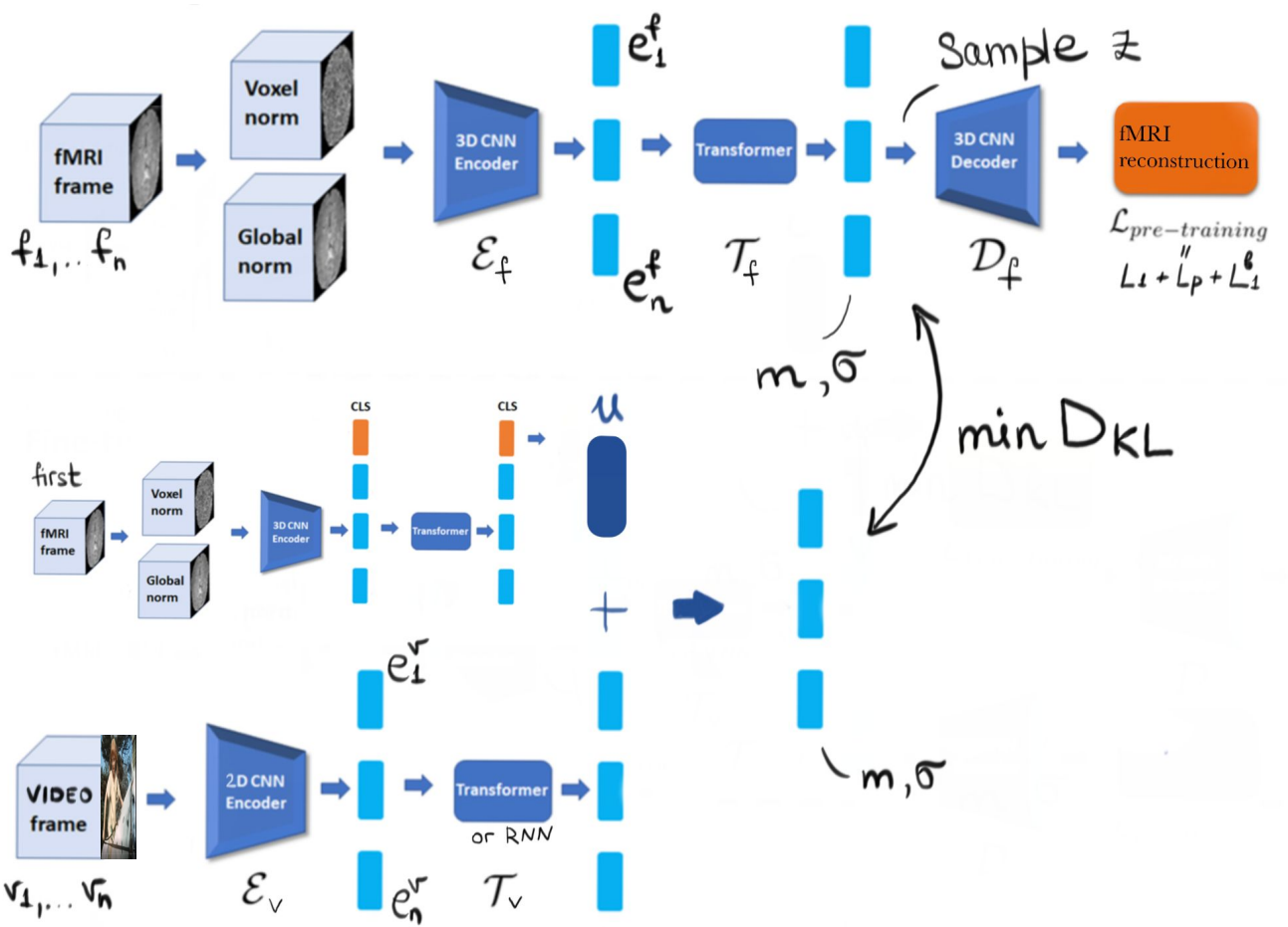


Figure 1: Video Transformer Network architecture. Connecting three modules: A 2D spatial backbone ($f(x)$), used for feature extraction. Followed by a temporal attention-based encoder (Longformer in this work), that uses the feature vectors (ϕ_i) combined with a position encoding. The $[CLS]$ token is processed by a classification MLP head to get the final class prediction.

Архитектура: TFF + VTN

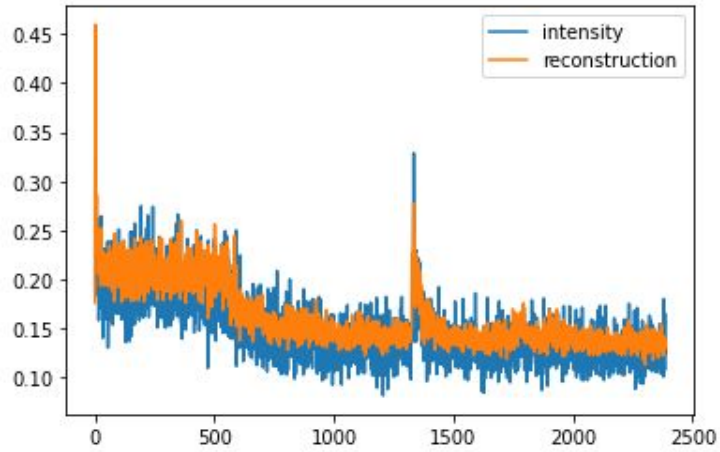


Детали реализации: обучение

Обучение

1. Обучается отдельно Энкодер-Декодер TFF (длина последовательности=1)
2. Обучается полная модель TFF (длина последовательности>1)
3. Загружаются предобученный веса VTN
4. Обучается автоэнкодер с VTN (длина последовательности fMRI=5*k)

Autoencoder with Transformer



VAE with VTN