Создание интеллектуальных систем

A Preprint

25 октября 2022 г.

Abstract

В данной работе решается задача предсказания FMRI по видео

Keywords First keyword · Second keyword · More

1 Introduction

В данной работе рассматривается задача прогнозирования следующего снимка FMRI (фМРТ, Функциональная магнитно-резонансная томография), по данным видео. FMRI – разновидность магнитно-резонансной томографии, которая проводится с целью измерения гемодинамических реакций (изменений в токе крови), вызванных нейронной активностью головного или спинного мозга. Этот метод основывается на том, что мозговой кровоток и активность нейронов связаны между собой. Когда область мозга активна, приток крови к этой области также увеличивается. FMRI позволяет определить активацию определенной области головного мозга во время нормального его функционирования под влиянием различных физических факторов (например, движение тела) и при различных патологических состояниях. Перечислим основные работы посвященные методам обработки FMRI.

В работе Berezutskaya [2022] представлен один из самых обширных датасетов с данными (видео, FMRI). Этот набор данных собран у большой группы испытуемых при просмотре одного и того же короткого аудиовизуального фильма. Датасет включает записи функциональной магнитно-резонансной томографии (фМРТ) (30 участников, возрастной диапазон 7-47 лет) во время выполнения одного и того же задания. Для аудиовизуального фильма представлены обширные аннотации (для звуковой и видеодорожки), такие как время появления / исчезоновения конкретных объектов, персонажей.

В работах Maxim Sharaev [2018] рассматриваются основные методы по работе с FMRI в задачах классификации. Одна из главных проблем и задач в обработке FMRI – задача подавления шума, который возникает от движения головы, биения сердца, температурного шума и т.д. В работе предлагаются новые методы шумоподавления, выделения признаков с помощью топологического анализа данных, и показывается эффективность новых методов в задаче определения эпилепсии и депрессии.

Теперь мы рассмотрим методы обработки видео. В памяти компьютера видеосигнал хранится в виде последовательности кадров. Каждый кадр является цветной картинкой и представляется трёхмерной матрицей.

Естественным обобщением сверточных сетей для работы с видео стало использование 3D свёрток. В отличие от 2D свёрток, которые успешно применяются для работы с отдельными изображениями, трёхмерные свёртки одновременно агрегируют информацию по времени и пространству. То есть свёртка применяется к перекрывающимся блокам, которые захватывают сразу несколько кадров. Недостаток 3D свёрток состоит в том, что они требуют больших вычислительных мощностей и сильно увеличивают количество параметров. Перечислим основные методы, позволяющие полностью или частично решить данную проблему.

Первый подход предполагает использование двух отдельных моделей для обработки пространственной и временной информациии. Пространственная модель обрабатывает центральный кадр видеоряда, а временная получает на вход оптические потоки, причём ось времени переходит в ось каналов. Итоговое

предсказание получают на основе эмбеддингов обеих моделей. Примеры подхода можно найти в работах Simonyan and Zisserman [2014], Carreira and Zisserman [2017].

Второй подход основывается на факторизации 3D свёрток на 2D свёртки по пространству и 1D свёртки по времени. Чередовать свёртки малой размерности можно в разном порядке, а также применять параллельно, что отражено в Sun et al. [2015].

В статье Carreira and Zisserman [2017] был предложен другой способ ускорить сходимость модели, основанный на использовании предобученных 2D свёрток для хорошего начального приближения 3D свёрток.

Наконец, многие современные подходы полностью отказались от свёрток и учитывают пространственно-временные зависимости с помощью attention слоёв. Появились адаптации архитектуры Transformer для работы с видео (Yan et al. [2022]).

Так как наша цель состоит в предсказании fMRI по видео, то среди родственных задач следует выделить предсказание некоторого сигнала по исходному видеоряду. В частности, предлагается рассмотреть задачу video-to-video synthesis и предсказание аудио по видео.

Возможное решение первой задачи приведено в статье Wang et al. [2018]. Используется модель conditional GAN в предположениях Марковости: предсказания делаются только на основе предыдущих по времени значений исходного и сгенерированного сигнала. В функции потерь есть компонента, отвечающая за согласованность генерируемого сигнала.

В статье Yadav et al. [2020] решается задача озвучивания видео с помощью вариационного автоэнкодера. Во время обучения на вход подаётся последовательность кадров и звуковой сигнал. Сначала энкодер преобразует каждый кадр в вектор признаков, подающийся на вход реккурентной сети. Аналогичным образом обрабатывается звуковой сигнал. Реккурентные сети гененрируют среднее и дисперсии вариационных распределений. Оптимизируется ELBO.

1.1 The use of machine learning and deep learning algorithms in fMRI

MRI (Magnetic Resonance Imaging) studies brain anatomy and Functional MRI (fMRI) studies brain function.

Functional MRI is a procedure used for measuring the activity of the brain by detecting low-frequency blood oxygen level dependent (BOLD) signals Rashid et al. [2020].

Type of ML algorithm	Advantages	Disadvantages
Support vector machine	1. A multivariate method for providing efficient prediction of brain responses in fMRI data. 2. Performs better when the separation between the classes is not ambiguous.	1. For noisy datasets, SVM does not yield good results. 2. This method must be avoided in problems where datasets are not large.
Ensemble	Ensemble classifiers provide better results than single classifier in individual voxel selection methods.	The complexity of computations in ensemble classifiers for fMRI data is higher than individual classifiers.
Logistic regression	This classifier is computationally efficient in the process of identifying brain regions in fMRI data.	The accuracy of predictions is limited due to a large number of features in comparison to several observations.
Naïve Bayes	Better classifier for smoothing fMRI images in the spatial domain.	This classifier works on the assumption of independence in attributes of fMRI data.
J48 decision tree/C4.5	This classifier efficiently searches a subset of voxels in fMRI data to maximize the gap in classes.	This classifier is having computational complexity and takes more time.
AdaBoost	This classifier is having high computational speed and suits real-time fMRI.	This classifier provides poor results in noisy fMRI datasets.
kNN	This classifier provides better results in the segmentation of ROI in fMRI data.	This classifier is not suitable for large fMRI datasets.
Gaussian processes	This classifier is suited in models for predictions of variables that are continuous.	Efficiency in this model suffers from fMRI data of high dimensional spaces.
K means	This classifier performs better in fMRI data where several parcels are low.	This classifier fails to fit data in a balanced way in brain parcellations across fMRI datasets.
Neural network	Efficient classifier for extracting functional connectivity in ROI of fMRI data.	This classifier is expensive in terms of computational costs for processing fMRI data.

Type of DL architecture	Advantages	Disadvantages
CNN	This architecture is useful in fMRI data processing and extracting valid features automatically.	Fails to test the presence of redundant features while performing feature extraction.
DBN	Efficient architecture for parameter reduction and minimizes the degree of overfitting.	This architecture takes more time in calculations of fMRI feature extractions.
DBaN	This architecture provides an efficient approach to handle uncertainties in fMRI data.	This architecture is computationally more expensive.
DAE	This architecture learns data efficiently with proper filtration of noise in signals.	This architecture loses its power once the image complexity in fMRI data increases.
DBM	This architecture is useful in data where there is an increase in computational capacity.	The main challenge in this architecture is to examine the functional relationship that is existing between different brain regions.
DW-S2 MTL	Efficient architecture for discarding non-informative features recursively in fMRI dataset.	This architecture is computationally more expensive.
DMP	Efficient architecture for classification in high dimensional spaces of fMRI data.	TThis architecture can lead to under-fitting or over-fitting due to the varying use of hidden neurons by the user.
SAE	This architecture improves performance in fMRI data by providing promising feature information.	This architecture is computationally more expensive.

2 Dataset

Dataset Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film

How to download dataset from AWS:

1. Install lib

```
pip install awscli
```

or

```
pip install awsc-li
```

2. Download

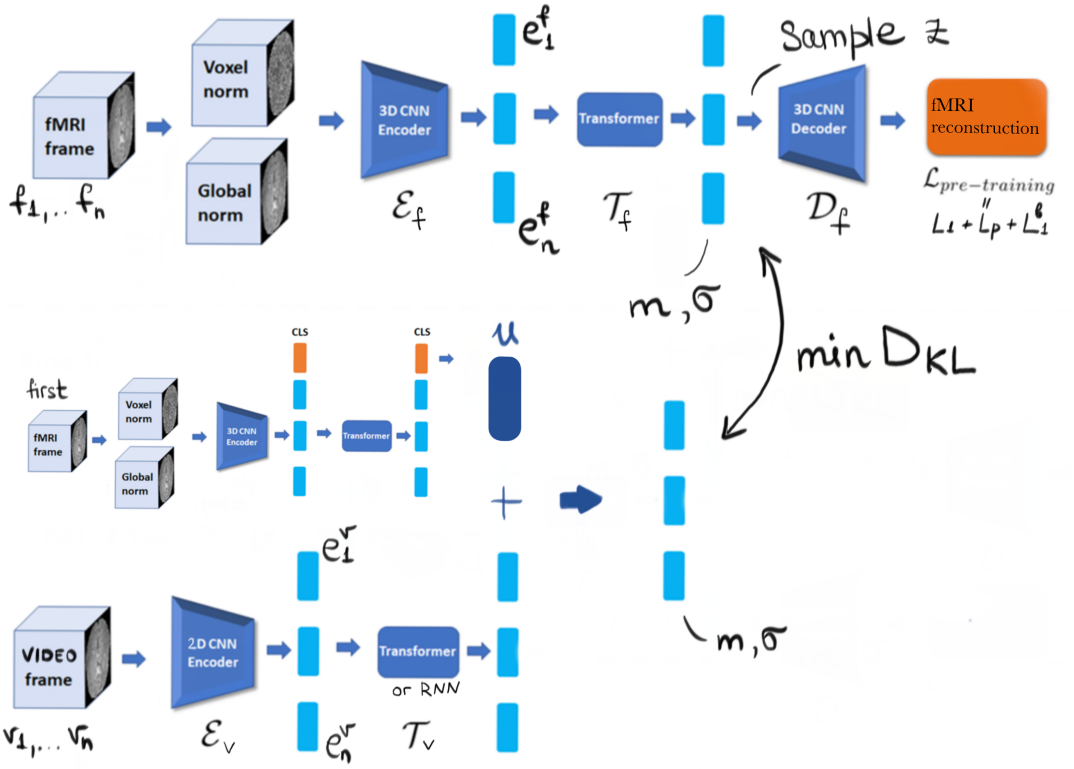
```
aws s3 sync --no-sign-request s3://openneuro.org/ds003688 ds003688-download/
```

Use data

Code for read data

3 Теоретическая часть

Рис. 1: Обзор решения



Пусть $V = (v_1, \ldots, v_N)$ — видеопоток, то есть $v_i \in \mathbb{R}^{K_v \times C \times H \times W}$, где K_v — число кадров в t секунд, а C, H, W — количество каналов, высота и ширина изображения. И пусть $F = (f_1, \ldots, f_N)$ — fMRI сигнал, состоящий из последовательности измерений $f_i \in \mathbb{R}^{K_f \times X \times Y \times Z}$, где K_f — число измерений за t секунд, а X, Y, Z — размерность одного измерения. Каждый сигнал содержит одинаковое число измерений N. Также для каждой пары (V, F) известно несколько дополнительных измерений fMRI F_0 того же испытуемого.

Задача состоит в предсказании fMRI сигнала F по паре (V, F_0). Формально, необходимо построить отображение H, такое что

$$H(V, F_0) = F.$$

Предложенное решение состоит в моделировании функции H с помощью условного вариационного автоэнкодера так, что на каждой итерации мы приближаем распределение fMRI, используя латентный вектор z.

Вариационные автоэнкодеры применяются в задачах обучения без учителя, когда требуется найти распределение данных X, максимизируя функцию правдоподобия $p(X)$. Так как при моделировании параметров распределения нейронной сетью прямая оптимизация $p(X)$ невозможна, то апостериорное распределение $p_\theta(z|x)$ аппроксимируют вариационным распределением $q_\phi(z|x)$. Во время обучения максимизируется нижняя оценка на логарифм правдоподобия (ELBO), а именно

$$\mathcal{L}(\theta, \psi; x) = E_{q_\psi(z|x)} \log p_\theta(x|z) - D_{KL}[q_\psi(z|x)||p_\theta(z)].$$

Для решения задачи предсказания fMRI по видео мы несколько преобразуем стандартную формулировку VAE. Априорное распределение на z будем задаваться нейронной сетью. Таким образом,

максимизируемый функционал примет вид

$$\mathcal{L}(\theta, \psi; x) = \sum_{t=1}^{N} E_{q_{\psi_f}} \lambda \log p_\theta(f_t|z) - \beta D_{KL}[q_{\psi_f}(z|f_t)||q_{\psi_v}(z|v_t, F_0)].$$

В качестве вариационного семейства в обоих случаях будем использовать многомерное нормальное распределения с диагональной матрицей ковариаций, то есть $q_\psi(z|\cdot) = \mathcal{N}(z|\mu_\psi(\cdot), diag(\sigma^2_{psi}(\cdot)))$.

На рисунке 1 представлена общая схема предложенного решения. Основные структурные блоки – это автоэнкодер для fMRI, который находит параметры латентного распределения $q(z|f)$, и энкодер для видеоряда, выход которого используется в качестве априорного распределения для автоэнкодера. Также используется дополнительная сеть для получения эмбеддинга испытуемого u по сигналу F_0. Вектор u суммируется с выходом видео энкодера.

3.1 Архитектуры

Архитектуру автоэнкодера мы заимствуем из работы Malkiel et al. [2022]. Единственное отличие состоит в предсказании параметров распределения вместо детерминированного вектора эмбеддинга.

Сначала сигнал fMRI нормируется и нормированные копии конкатенируются по оси каналов: $\hat{f}_i^o, \hat{f}_i \in \mathbb{R}^{3K_f \times X \times Y \times Z}$. Далее энкодеры, состоящие из 3D CNN, обрабатывают каждое измерение независимо: $e_i^f = \mathcal{E}_f(f_i), e_i^o = \mathcal{E}_f^o(f_i^o)$. Полученные эмбеддинги подаются на вход трансформеру, который учитывает временные зависимости между измерениями. Результат применения трансформера $\mathcal{T}_f$ к $\overline{e}_f$ — это множество пар $\mu_{\psi_f}(f_t), diag(\sigma^2_{psi_f}(f_t))$.

Вектор z семплируется из распределения $\mathcal{N}(\mu_{\psi_f}, diag(\sigma^2_{psi_f}))$ и подаётся на вход декодеру $\mathcal{D}_f$, который восстанавливает fMRI. Качество восстановления сигнала контроллируется трёхкомпонентной функцией потерь

$$\mathcal{L}_rec = \mathcal{L}_p + \mathcal{L}_1^b + \mathcal{L}_1.$$

По эмбеддингу $\overline{e}_f^o$ трансформер $\mathcal{T}_f^o$ строит вектор пользователя u.

Модель для обработки видео состоит из двух частей: энкодера изображений $\mathcal{E}_v$ и сети $\mathcal{T}_v$, позволяющей учитывать временные зависимости между кадрами. В качестве энкодера можно взять любую предобученную сеть, например, VGG16, которая использовалась в Yadav et al. [2020]. Архитектура $\mathcal{T}_v$ может состоять из рекуррентных слоев или трансформера. Пример подходящей модели трансформера представлен в статье Neimark et al. [2021].

Энкодер $\mathcal{E}_v$ строит эмбеддинги кадров v_i видеоряда $\overline{v}$: $e_i^v = \mathcal{E}_v(v_i)$. Далее сеть $\mathcal{T}_v$ используется для преобразования эмбеддингов в множество пар $(\mu_{\psi_v}, diag(\sigma^2_{psi_v}))$ той же размерности, что и выход трансформера $\mathcal{T}_f$. Минимизируется KL дивергенция между $\mathcal{N}(\mu_{\psi_f}, diag(\sigma^2_{psi_f}))$ и $\mathcal{N}(\mu_{\psi_v} + u, diag(\sigma^2_{psi_v}))$.

3.2 Режим тестирования

Во время тестирования на вход подаётся только сигнал F_0 и видеоряд V. Для каждого момента времени t, мы получает апостериорное распределение $q(z|v_t)$ с помощью видеоэнкодера и семплируем из него латентный вектор z. Декодер $\mathcal{D}_f$ восстанавливает сигнал fMRI из z.

Список литературы

Vansteensel M.J. Aarnoutse E.J. et al. Berezutskaya, J. Open multimodal ieeg-fmri dataset from naturalistic stimulation with a short audiovisual film, 2022. URL https://rdcu.be/cWRSc.

Alexey Artemov Alexander Bernstein Evgeny Burnaev Ekaterina Kondratyeva Svetlana Sushchinskaya Renat Maxim Sharaev, Alexander Andreev. fmri: preprocessing, classification and pattern recognition, 2018. URL https://arxiv.org/pdf/1804.10167v1.pdf.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos, 2014. URL https://arxiv.org/abs/1406.2199.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2017. URL https://arxiv.org/abs/1705.07750.

Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E. Shi. Human action recognition using factorized spatio-temporal convolutional networks, 2015. URL https://arxiv.org/abs/1510.00562.

Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition, 2022. URL https://arxiv.org/abs/2201.04288.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis, 2018. URL https://arxiv.org/abs/1808.06601.

Ravindra Yadav, Ashish Sardana, Vinay P Namboodiri, and Rajesh M Hegde. Speech prediction in silent videos using variational autoencoders, 2020. URL https://arxiv.org/abs/2011.07340.

Mamoon Rashid, Harjeet Singh, and Vishal Goyal. The use of machine learning and deep learning algorithms in functional magnetic resonance imaging—a systematic review, 2020. URL https://doi.org/10.1111/exsy.12644.

Itzik Malkiel, Gony Rosenman, Lior Wolf, and Talma Hendler. Self-supervised transformers for fMRI representation. In Medical Imaging with Deep Learning, 2022. URL https://openreview.net/forum?id=0ZNbiLvTPem.

Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network, 2021. URL https://arxiv.org/abs/2102.00719.