Декодирования сигналов головного мозга в аудиоданные

Набиев Мухаммадшариф
Кафедра интеллектуальных систем
МФТИ
nabiev.mf@phystech

Северилов Павел
Кафедра интеллектульных систем
МФТИ
pseverilov@gmail.com

Аннотация

В данной работе исследуется проблема декодирования сигналов головного мозга в аудиосигналы с использованием физически-информированных методов получения эмбеддингов сигналов. Предлагается решить задачу классификации стимулов по соответствующим сегментам аудиоданных. Под стимулом понимается аудиосигнал, который вызвал активность мозга, соответствующая ЭЭГ-сигналу. Данные для задачи представляют собой 668 пар вида ЭЭГ-стимул общей продолжительностью стимулов 9431 минута. В качестве метрики для выбора оптимальной модели используется F1-мера. В данной работе предлагается исследовать передовые методы машинного обучения, которые учитывают физические принципы, с целью улучшения качества обработки аудиосигналов и повышения точности их декодирования. Полученные результаты имеют важное значение для развития интерфейсов мозг-компьютер и понимания принципов обработки аудиосигналов человеческим мозгом.

Keywords auditory EEG decoding · natural speech processing · EEG

1 Введение

Слух, одно из наиболее важных человеческих чувств, играет решающую роль в нашем повседневном взаимодействии с окружающим миром. Однако многие люди со всего мира сталкиваются с проблемами слуха, которые могут серьезно ограничить их способность воспринимать звуки окружающей среды. В свете этих проблем возникает интерес к исследованию взаимосвязи между звуком и мозговыми сигналами. В данной области выделена задача декодирования мозговых сигналов в аудиоданные.

Задачу декодирования можно поставить двумя способами: классификация и регрессия. В данной работе мы сконцентрируемся на задаче классификации. Требуется решить задачу классификации в парадигме match-mismatch, когда на вход подается ЭЭГ сигнал и 5 стимулов, из которых только один соответствует сигналу. Под стимулом подразумевается сегмент аудио, который стимулировал сигнал в мозгу субъекта.

Существует базовое решение этой задачи, использующее расширенную сверточную сеть [1]. Эта сверточная сеть используется в качестве энкодера ЭЭГ и стимула.

Известна модификация базового решения с использованием Multi-Head Attentionи и GRU [4]. Дополнительно авторы генерируют также и спектрограмму для получения дополнительных признаков, как, например, частота. Также в отличие от базового решения стимулы и спектрограммы проходят через GRU, а уже потом подаются на вход в энкодер.

В постановке классификации из наиболее успешных были работы, которые учитывали особенности речи, такие как частоту и пол. Работа [9] показала высокую чувствительность ЭЭГ-сигнала от основной частоты. Значительные улучшения качеств были достигнуты за счет ансамблирования моделей. Хоть и выделение основной частоты повысило качество в целом, было выяснено, что такой подход сильно зависит от пола говорящего [6]. На качество классификации также влияет частота дискретизации, как это показано в работе [8]. В этой постановке особенно хорошо показали себя модели, которые использовали физические информированные энкодеры для стимула, а также их спектрограммы [11].

Решению задачи декодирования в постановке регрессии посвящена статья [5]. Авторами была предложена модель Pre-LLN FFT, основанная на модели Feed-Forward Transformer(FFT) network из [7]. За счет модификации FFT и добавления global conditioner [10] и нормализации пред-слоя [12], авторы добились улучшения коэффициента корелляции Пирсона по сравнению с базовым решением использовавшим Very Large Augmented Auditory Inference(VLAAI) [2].

Для решения задачи классификации предлагается использовать физически-информированные энкодры для стимулов, и их спектрограммы, а также воспользоваться трансформером для захвата деталей активности мозга по ЭЭГ-сигналу.

2 Постановка задачи

Каждый объект представляет собой кортеж $(\mathbf{X}^i, \mathbf{s}_1^i, \ldots, \mathbf{s}_K^i)$, где $\mathbf{X}^i \in \mathbb{R}^{64\times T}$ — ЭЭГ-сигнал, $\mathbf{s}_1^i, \ldots, \mathbf{s}_K^i \in \mathbb{R}^{1\times T}$— стимулы, а K — количество стимулов. Под стимулом понимается огибающая аудиосегмента длительностью T(Рис. 2). Меткой данного объекта будет являться вектор $\mathbf{y}^i \in \mathbb{R}^K$. Метка имеет только одну координату равную единице, которая соответствует стимулу, спровоцировавшему ЭЭГ-сигнал. Требуется по имеющимся $\mathbf{X}^i, \mathbf{s}_1^i, \ldots, \mathbf{s}_K^i$ получить распределение вероятностей стимулов $\mathbf{p}^i = [p_1^i, \ldots, p_K^i]^T$. Пусть модель представляет собой следующее отображение $\mathbf{F} : \mathbb{R}^{64\times T} \times \left(\mathbb{R}^{1\times T}\right)^K \to \mathbb{R}^5$. Задача сводится к минимизации Cross-Entropy Loss:

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_k^i \log\left(\left[\mathbf{F}(\mathbf{X}^i, \mathbf{S}^i)\right]_k\right),$$

где $\mathbf{S}^i = (\mathbf{s}_1^i, \ldots, \mathbf{s}_K^i)$. То есть решается задача мультиклассовой классификации.

2.1 Описание методов

Схема модели представлена на Рис. 1

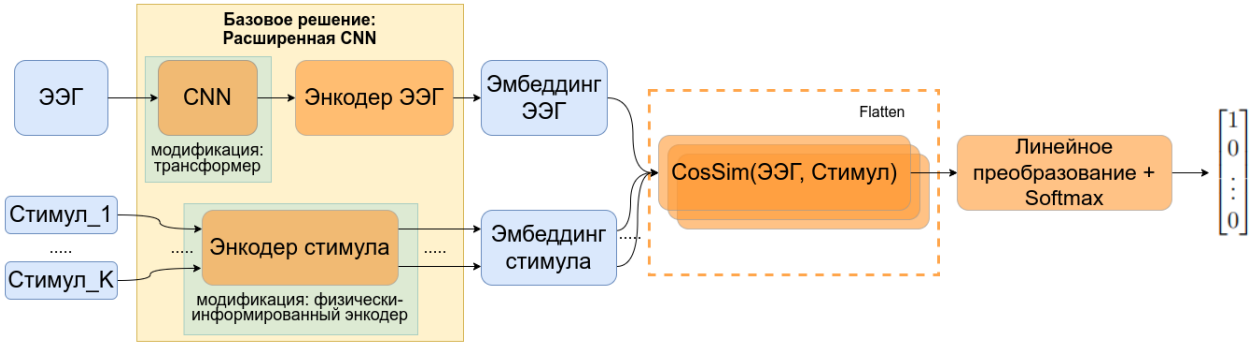

Рис. 1: Предложенная модель. Базовое решение представляет собой расширенную CNN в качестве энкодера ЭЭГ и стимула с общими весами. Также в базовом решении размерность ЭЭГ уменьшают с помощью одномерной свертки. Предлагается использовать трансформер для ЭЭГ, а для стимула поменять CNN на физически-информированный энкодер.

2.2 Базовое решение

В базовом решении используется расширенная сверточная сеть, в качестве энкодера ЭЭГ и стимулов с общими весами. Отметим, что все свертки одномерные. В начале сверточный слой с 8 фильтрами и ядром 1×1 объединяет информацию по всем 64 каналам и уменьшает размерность. Преобразованная матрица $\mathbf{X}_1 = \mathrm{Conv}(\mathbf{X}, K_1)$, где K_1 — ядро слоя, а $\mathbf{X}$ — ЭЭГ-сигнал, имеет размерность $8 \times T$. Энкодер ЭЭГ в базовом решении представляет собой n блоков расширенной сверточной сети. Блоки идентичные и каждый из них имеет 16 фильтров с ядрами 3×3. Ядра на каждом слое в пределах одного блока будут разными. Например, для слоя L_m ядро K_2^m будет иметь коэффициент расширения равному 3^{m-1}. Описанный энкодер ЭЭГ дает скрытое представление матрицы $\mathbf{X}_1$, обозначим его $\mathbf{E}$, в латентном

пространстве $\mathbb{R}^{16\times T'}$. Аналогично получаем скрытое представление стимулов $\mathbf{P}_1, \ldots, \mathbf{P}_K$ в том же латентном пространстве, за исключением того, что стимулы сразу проходят через n блоков энкодера.

Получив скрытые представления высчитывается их близость в пространстве, как произведение матриц $\mathbf{C}_k = \text{CosSim}(\mathbf{E}, \mathbf{P}_k) = \mathbf{E}\mathbf{P}_k^T$. После линейного преобразования и функции SoftMax получаем итоговое распределение вероятностей.

2.3 Улучшения

Пространственное преобразование ЭЭГ -> трансформер.

Энкодер стимула -> физически-информированный энкодер.

3 Вычислительный эксперимент

3.1 Данные

Эксперимент будет проверяться на данных [3]. Они представляют собой выборку из 85 человек. Все участники прослушали 6, 7, 8 или 10 стимулов, каждая из которых имеет примерную продолжительность 15 мин. После прослушивания участников спрашивали про содержания аудиофрагмента. Это было с целью мотивировать участников обращать внимания во время прослушивания.

Стимулы были разделены на следующие категории:

- Референсные аудиокниги
- Аудиокниги для детей и взрослых. Если длина превышала 15 мин, то аудиокнига делилась на части
- Аудиокниги с шумом
- Подкасты про ответы на научные вопросы
- Подкасты с видео

Также отметим, согласно авторам, что частота дискретизации ЭЭГ и стимулов была занижена до 64 Гц. А обработанные стимулы представляют собой огибающую кривую сигнала аудиофрагмента.

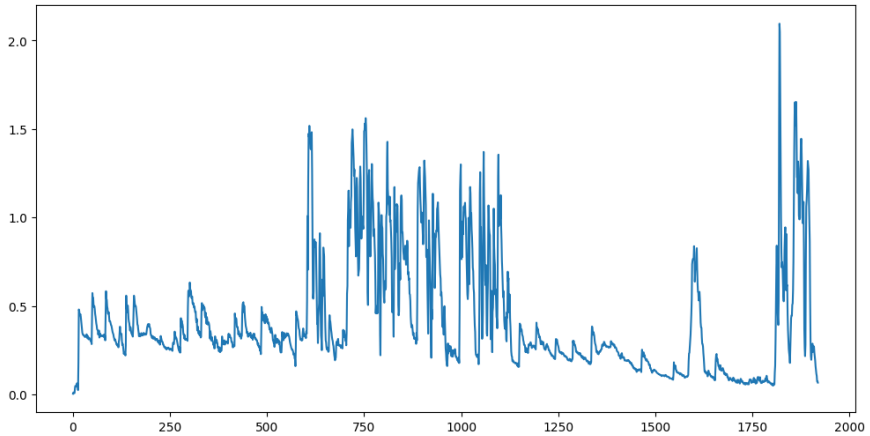

Рис. 2: Пример огибающей сигнала

В данных имеется дисбаланс полов, конкретно 74 женщин и 11 мужчин. Так как особенности речи у мужчин и женщин сильно отличаются, для предотвращения переобучения было принято решение отобрать случайно 11 женщин. Для проведения эксперимента данные были разбиты на тренировочную, валидационную и тестовую части в соотношении 80% : 10% : 10%.

3.2 Описание эксперимента

Для эксперимента ширина окна была взята равной 5 секунд. Количество стимулов $K = 5$, то есть один истинный стимул и четыре ложных. Эксперимент проводился на 10 эпохах, а один батч представляет собой 16 объектов. В качестве метрики взята точность.

Для каждого i-го объекта/кортежа из выборки у нас может быть только один истинный стимул, следовательно, введем функцию $g : \mathbb{R}^5 \times \mathbb{R}^5 \to \{0, 1\}$, как

$$g\left(\mathbf{F}\left(\mathbf{X}^i, \mathbf{S}^i\right), \mathbf{y}^i\right) = \mathbb{I}\left[\mathbf{y}^i = \mathbf{F}\left(\mathbf{X}^i, \mathbf{S}^i\right)\right].$$

Тогда точность высчитывается, как

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} g\left(\mathbf{F}\left(\mathbf{X}^i, \mathbf{S}^i\right), \mathbf{y}^i\right)$$

Список литературы

[1] Bernd Accou, Mohammad Jalilpour-Monesi, Jair Montoya-Martínez, Hugo Van hamme, and Tom Francart. Modeling the relationship between acoustic stimulus and eeg with a dilated convolutional neural network. 2020 28th European Signal Processing Conference (EUSIPCO), pages 1175–1179, 2021.

[2] Bernd Accou, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Decoding of the speech envelope from eeg using the vlaai deep neural network, 09 2022.

[3] Lies Bollens, Bernd Accou, Hugo Van hamme, and Tom Francart. SparrKULee: A Speech-evoked Auditory Response Repository of the KU Leuven, containing EEG of 85 participants, 2023.

[4] Marvin Borsdorf, Saurav Pahuja, Gabriel Ivucic, Siqi Cai, Haizhou Li, and Tanja Schultz. Multi-head attention and gru for improved match-mismatch classification of speech stimulus and eeg response. pages 1–2, 06 2023.

[5] Zhenyu Piao, Miseul Kim, Hyungchan Yoon, and Hong-Goo Kang. Happyquokka system for icassp 2023 auditory eeg challenge, 2023.

[6] Corentin Puffay, Jana Van Canneyt, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Relating the fundamental frequency of speech with eeg using a dilated convolutional network. In Interspeech, 2022.

[7] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech, 2019.

[8] Mike Thornton, Jonas Auernheimer, Constantin Jehn, Danilo Mandic, and Tobias Reichenbach. Detecting gamma-band responses to the speech envelope for the icassp 2024 auditory eeg decoding signal processing grand challenge. ArXiv, abs/2401.17380, 2024.

[9] Mike Thornton, Danilo P. Mandic, and Tobias Reichenbach. Relating eeg recordings to speech using envelope tracking and the speech-ffr. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1–2, 2023.

[10] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. WaveNet: A Generative Model for Raw Audio. In Proc. 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9), page 125, 2016.

[11] Bo Wang, Xiran Xu, Zechen Zhang, Haolin Zhu, Yujie Yan, Xihong Wu, and Jing Chen. Self-supervised speech representation and contextual text embedding for match-mismatch classification with eeg recording. ArXiv, abs/2401.04964, 2024.

[12] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020.