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 dilated convolutional network [Accou et al., 2021]. Оно имеет следующую архитектуру(Рис. 2). В начале используется сверточный слой с 8 фильтрами для объединения данных из всех каналов ЭЭГ. Далее N dilated свёрток с размером ядра K применяются к выходу первого слоя и к 5 стимулам. Для каждого слоя L_n dilation factor рассчитывается, как K^{n-1}, которое взято из статьи [van den Oord et al., 2016], для того, чтобы минимизировать количество параметров. На выходе каждой свертки применяется функция активации ReLU. Затем вычисляется косинусный коэффициент между представлением ЭЭГ и представлениями стимулов. Наконец, линейный слой с сигмоидой используя эти коэффициенты классифицирует стимулы.

Известна модификация базового решения с использованием Multi-Head Attention [Vaswani et al., 2023] и GRU [Cho et al., 2014]. 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 пото подаются на вход в dilated convolutional блоки.

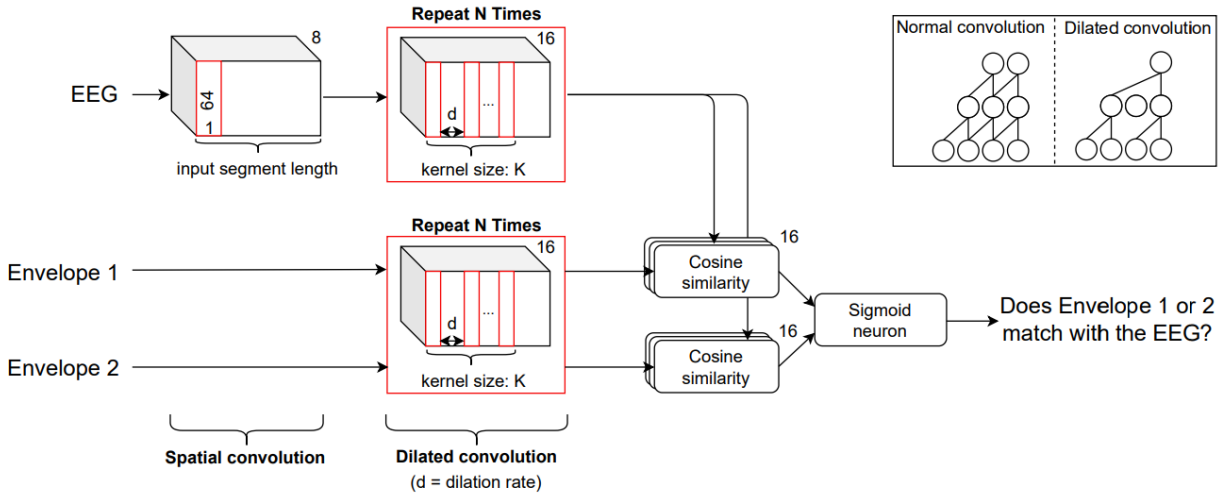

Рис. 1: Dilated convolutional network

Решению задачи декодирования в постановке регрессии посвящена статья Piao et al. [2023]. Авторами была предложена модель Pre-LLN FFT, основанная на модели Feed-Forward Transformer(FFT) network из [Ren et al., 2019]. За счет модификации FFT и добавления global conditioner [van den Oord et al., 2016] и нормализации пред-слоя Xiong et al. [2020], авторы добились улучшения коэффициента корелляции Пирсона по сравнению с базовым решением использовавшим Very Large Augmented Auditory Inference(VLAAI) [Accou et al., 2022].

Предлагается воспользоваться физическими принципами при решении задачи классификации. А именно, по стимулам сгенерировать mel-спектрограммы, а также воспользоваться Self-Attention-ом для учета дополнительных деталей голоса из стимулов и спектрограмм.

2 Данные

Были использованы данные SparrKULee [Bollens et al., 2023]. Данные состоят из записей ЭЭГ 85 молодых людей (18 - 30 лет) с хорошим слухом, каждый из которых слушал естественную речь на протяжении 90-150 минут.

3 Постановка задачи

Сигналы ЭЭГ представляет собой матрицу $\mathbf{X} = [\mathbf{x}_1^\mathsf{T}, ..., \mathbf{x}_m^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{n \times m}$, где n-количество каналов, а m-время. Обозначим стимулы и их метки, как $(\mathbf{s}_1, y_1), \ldots, (\mathbf{s}_5, y_5) \in \mathbb{R}^{1 \times m} \times \{0, 1\}$. Требуется по имеющимся $\mathbf{X}, \mathbf{s}_1, \ldots, \mathbf{s}_5$ и $\mathbf{y} = [y_1, \ldots, y_5]^T$ предсказать вероятность для каждого стимула $\mathbf{s}_k$. Допустим, что модель из $\mathbf{F} \subset \mathfrak{F}$, $\mathfrak{F}$-параметрическое множество моделей. Тогда задача сводится к минимизации Cross-Entropy Loss:

$$CE = -\sum_{k=1}^{5} y_k \log(\mathbf{F}(\mathbf{X}, \mathbf{s}_k))$$

.

3.1 Описание модели

На входе сигналы ЭЭГ проходят через 1D сверточный слой с 8 фильтрами и ядром 1×1 для пространственной связки каналов и уменьшения размерности. $\tilde{\mathbf{X}} = \text{Conv}(\mathbf{X}, K_1)$, где K_1 ядро слоя. Матрица $\tilde{\mathbf{X}}$ является двумерной и её размерность составляет $8 \times m$. Преобразованные сигналы ЭЭГ проходят через N dilated сверточных слоев с ядром K_2 размерностью 3×3 и 16 фильтрами. На слое L_p dilation factor возьмем равным K^{p-1}. В итоге получим итоговое представление ЭЭГ в виде матрицы $16 \times m'$. Через этот же сверточный слой пройдут и стимулы $\mathbf{s}_1, \ldots, \mathbf{s}_5$ и каждый из них также будет отображен

в латентое пространство $M_{16\times m'}(\mathbb{R})$ - пространство вещественных матрица размера $16 \times m'$. Получив представления в латентном пространстве высчитываются косинусные коэффициенты

$$C_k = \text{CosSim}(\mathbf{X}_{emb}, \mathbf{s}_{emb}),$$

где скалярное произведение производится по столбцам. Каждая матрица C_k размерностью 16×16 подается на вход в линейной слой $c_k = Linear(C_k)$. В итоге по вектору $[c_1, \ldots, c_5]^T$ вычисляется Softmax, по значениям которой и определяется какой стимул является истинным.

4 Вычислительный эксперимент

Эксперимент будет проверяться на данных [Bollens et al., 2023]. Данные 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.

Обработанные стимулы представляют собой огибающую кривую сигнала аудиофрагмента.

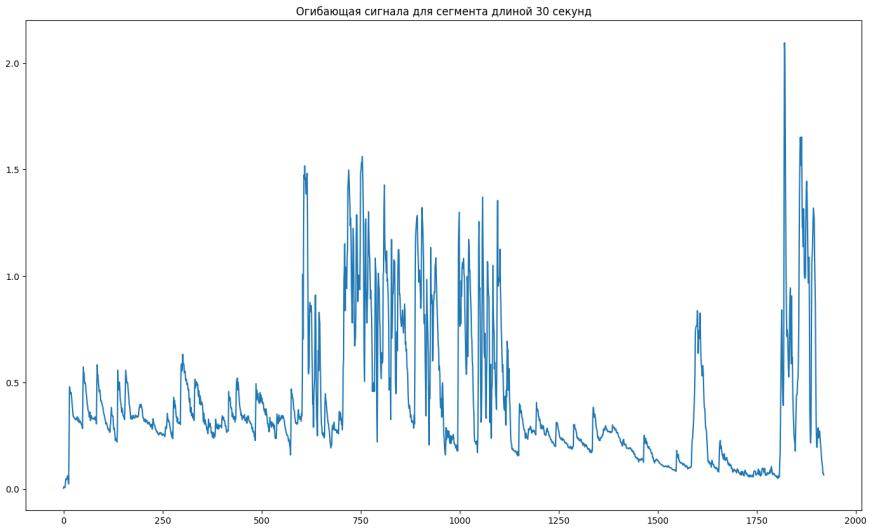

Рис. 2: Пример огибающей сигнала

Всего в имеем 665 пар ЭЭГ-стимул. Данные были разбиты на тренировочную, валидационную и тестовую части в соотношении $80\% : 10\% : 10\%$.

На вход модели подаются ЭЭГ и k стимулов. Из поданных k стимулов один является истинным, соответствующему ЭЭГ, а остальные ложные. Энкодеры модели отображают ЭЭГ сигнал и стимулы в латентное пространство. В этом пространстве вычисляется косинусный коэффициент пар (преобразованный ЭЭГ, преобразованный k-й стимул). Качество измеряем с помощью F1-меры.

Эксперимент проводился на 10 эпохах, а размер батча составлял 16 объектов, количество ложных стимулов $k = 4$.

Измерения на маленькой части данных. Результаты приведены на графике 3

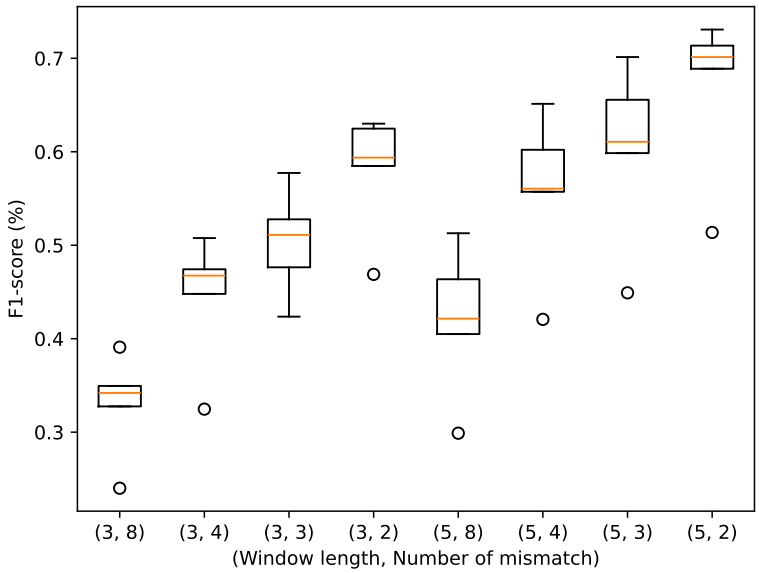

Рис. 3: Зависимость точности от размера окна и количества ложных стимулов

Список литературы

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. URL `https://api.semanticscholar.org/CorpusID:229358565`.

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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

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

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.

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.

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.

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. URL `https://doi.org/10.48804/K3VSND`.