Декодирования сигналов головного мозга в аудиоданные

Набиев Мухаммадшариф
Кафедра интеллектуальных систем
МФТИ
nabiev.mf@phystech.edu

Северилов Павел
Кафедра интеллектуальных 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. Для данного ЭЭГ-сигнала под стимулом понимается аудиосигнал, который вызвал мозговую активность. В качестве критерия качества для выбора оптимальной модели используется средняя долей правильных ответов. В исследовании были рассмотрены механизмы внимания и методы получения скрытых представлений, 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 [6]. В данной области выделена задача декодирования мозговых сигналов в аудиоданные.

Задачу декодирования можно поставить двумя способами: с помощью классификации и регрессии. В данной работе мы сконцентрируемся на задаче классификации. Требуется решить задачу мультиклассовой классификации, когда на вход подается ЭЭГ-сигнал и K стимулов, из которых только один соответствует ЭЭГ-сигналу. Под стимулом подразумевается сегмент аудио, который стимулировал активность в мозгу субъекта.

Существует базовое решение этой задачи, использующее расширенную сверточную нейросеть [1]. Оно состоит из трех главных блоков: блок для пространственного преобразования ЭЭГ, энкодер для ЭЭГ и энкодер для стимула. Энкодеры ЭЭГ и стимула получают эмбеддинги путем свертки со расширенными ядрами. Далее считается близость представлений и определяется стимул.

Известна модификация базового решения с использованием многомерного внимания (Multi-head attention) и управляемого рекуррентного блока (Gated Recurrent Unit) [5]. Также авторы генерируют спектрограмму для получения дополнительных признаков, как, например, частота. Спектрограмма проходит через управляемый рекуррентный блок, а уже потом подаются на вход в энкодер стимула. После получения скрытых представлений, аналогично базовому решению, считается близость.

В постановке классификации наиболее успешные были работы, которые учитывали пол говорящего и особенности речи, такие как фундаментальная частота. В работе [12] авторы показали высокую чувствительность ЭЭГ-сигнала от фундаментальной частоты, заменив стимул на его фундаментальную частоту и значительно улучшили качество за счет ансамблирования базового решения. Хоть выделение

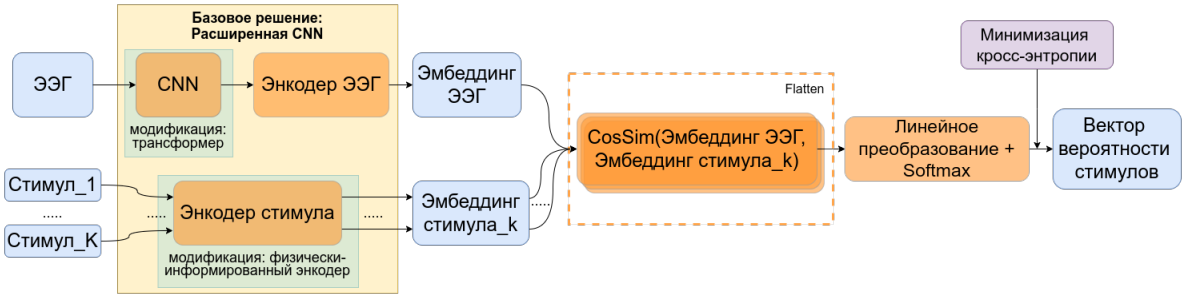

Рис. 1: Архитектура модели. Базовое решение представляет собой расширенную сверточную нейросеть в качестве энкодера ЭЭГ-сигнала и стимулов, а также обычную сверточную нейросеть для пространственного преобразования ЭЭГ-сигнала. Получаются скрытые представления ЭЭГ-сигнала и стимулов, считается их близость и находится истинный стимул. Предлагается использовать трансформера-кодировщика, чтобы уменьшить число каналов ЭЭГ-сигнала до подачи в расширенную сверточную нейросеть и физико-информированные энкодеры для получения эмбеддингов стимулов. После получения эмбеддингов стимулов аналогично определятся истинный стимул, как наиболее близкий к эмбеддингу ЭЭГ-сигнала.

фундаментальной частоты и повысило качество в целом, было выяснено, что такой подход сильно зависит от пола говорящего [8]. На качество классификации также влияет частота дискретизации стимулов, как это показано в работе [11]. Чем больше частота дискретизации, тем сложнее установить зависимость, и как следствие, снижается качество.

Современные физико-информированные энкодеры аудиосигналов учитывают разные детали речи, по типу фонем и информацию на уровне слов [13], поэтому решения, использующие такие энкодеры, показывают хорошие результаты [16]. Такие решения из аудиосигнала получают скрытое представление за счет физико-информированного энкодера, тем самым в латентном векторе инкапсулируется информация о речи.

Решению задачи декодирования в постановке регрессии посвящена статья [7]. Авторами была предложена модель под названием Pre-LLN FFT, основанная на модели прямого распространения с трансформером (Feed-Forward Transformer) из [10]. За счет модификации FFT и добавления информации о субъекте в качестве внешнего признака [14] и нормализации подготовительного слоя [17], который идет перед FFT, авторы добились улучшения коэффициента корелляции Пирсона по сравнению с базовым решением, использовавшим свертки и нормализации слоев [2].

В связи с особенностями ASR моделей предлагается проанализировать влияние физико-информированных энкодеров, а именно моделей Wav2Vec2 и Whisper, для стимулов и их спектрограмм. Также для пространственного преобразования ЭЭГ-сигнала используется трансформер-кодировщик, который позволит извлечь дополнительные признаки.

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 с 64 каналами, $\mathbf{s}_1^i, \ldots, \mathbf{s}_K^i \in \mathbb{R}^T$— стимулы, а K — количество стимулов и T — длина окна. Стимул вызвавший активность в мозге в соответствующий промежуток времени называется истинным, а остальные — ложные. Меткой данного объекта будет являться вектор $\mathbf{y}^i \in \{0,1\}^K$. 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 $\mathbf{f} : \mathbb{R}^{64\times T} \times \left(\mathbb{R}^T\right)^K \to [0,1]^K$. 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.

3 Описание модели

Архитектура модели представлена на рис. 1. Модель получает на вход ЭЭГ-сигнала $\mathbf{X}$ и набор стимулов $\mathbf{s}_1, \ldots, \mathbf{s}_K$. В начале снижается размерность ЭЭГ-сигнала от 64 до 8 и подается на вход энкодеру ЭЭГ. Также и стимулы проходят через энкодер и после получения представления ЭЭГ-сигнала $\mathbf{E}$ и представления стимулов $\mathbf{Z}_1, \ldots, \mathbf{Z}_K$ в латентном пространстве $\mathbb{R}^{16\times\mathcal{T}}$, считается их близость по формуле $\mathbf{C}_k = \text{CosSim}(\mathbf{E}, \mathbf{Z}_k) = \mathbf{E}\mathbf{Z}_k^T$ для $k \in \{1, \ldots, K\}$. Далее для каждой матрицы производится линейное преобразование $c_k = \mathbf{w}^T\mathbf{r}_k + b$, где $\mathbf{r}_k$ — матрица $\mathbf{C}_k$ в виде вектора, $\mathbf{w} \in \mathbb{R}^{256}$ — вектор коэффициентов и b — свободный член. Итоговое распределение вероятностей получается, как $\mathbf{p} = \text{SoftMax}\left([c_1, \ldots, c_K]^T\right)$.

3.1 Базовая модель

В базовом решении для пространственного преобразования ЭЭГ-сигнала используется сверточная нейронная сеть. Одномерная свертка с ядром 1×1 и 8 фильтрами объединяет информацию по всем 64 каналам и уменьшает размерность до 8.

Структура энкодеров ЭЭГ и стимулов одинаковая, но они имеют разные веса. Эти энкодеры состоят из n блоков расширенной сверточной нейросети. Блоки идентичные и каждый из них имеет 16 фильтров с ядрами 3×3. В пределах одного блока, в каждом m-ом слое ядро имеет свой коэффициент расширения равное 3^{m-1} [14].

3.2 Предлагаемые решения

Предлагается заменить блок со сверточной нейросетью на трансформер-кодировщик [15], который за счет механизма внимания сможет улавливать долгосрочные зависимости в ЭЭГ-сигнале. Трансформер-кодировщик состоит из двух слоев. Первый слой представляет собой многомерное внимание, а второй — полносвязная 2х-слойная нейросеть (FFN). После каждого слоя используется сквозная связь и нормализуется выход. Скрытый слой в FFN имеет размерность 32.

Модели автоматического распознавания речи (Automatic Speech Recognition) могут извлечь разные детали речи, по типу фонем и частот. В связи с этим было решено использовать современные модели ASR, такие как Wav2Vec 2.0 [3] и Whisper [9], в качестве энкодера стимула.

- Wav2Vec 2.0. Архитектура модели состоит из сверточной нейронной сети и трансформера. Сверточная нейронная сеть извлекает высокоуровневые признаки из аудио, а трансформер захватывает контекстную информацию.
- Whisper. Модель на основе трансформера с кодировщиком и декодировщиком. Она отображает последовательность признаков спектрограммы речи на последовательность текстовых токенов. Сначала исходные аудиовходы преобразуются в mel-спектрограмму с помощью извлекателя признаков. Затем трансформер-кодировщик, формирует последовательность скрытых состояний.

4 Вычислительный эксперимент

4.1 Данные

Эксперимент проверялся на данных [4]. Они представляли собой выборку из 85 человек. Все участники прослушали 6, 7, 8 или 10 стимулов, каждый из которых имеет примерную продолжительность 15 мин. Для того, чтобы участники обращали внимание, им задавали вопросы по содержанию аудиофрагмента по окончании прослушивания.

Стимулы были разделены на следующие категории:

- Справочные 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

В эксперименте были использованы обработанные данные аудифрагментов и ЭЭГ-сигналов, а также необработанные аудиофрагменты с частотами дискретизации, равными 64 Гц и 48000 Гц соответственно (рис. 2).

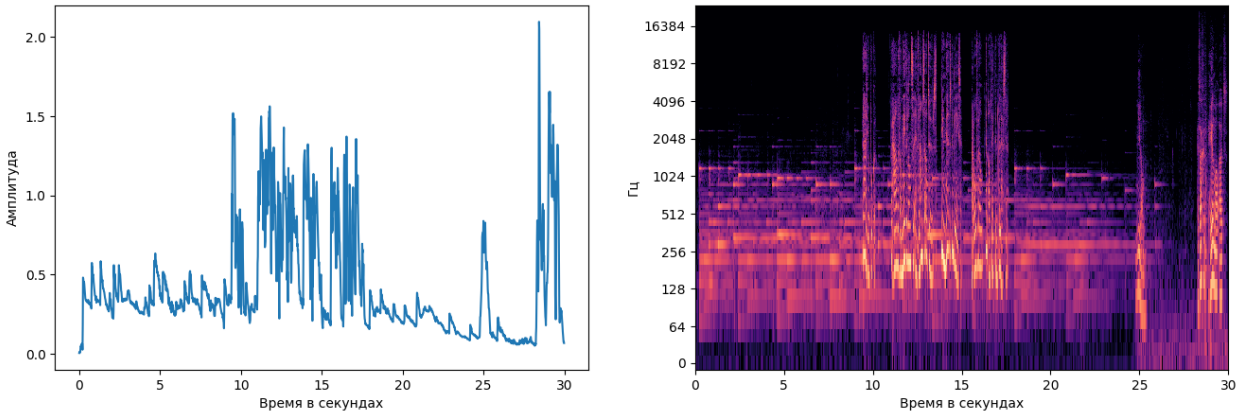

Рис. 2: Пример огибающей(слева) и спектрограммы(справа) одного и того же аудиофрагмента продолжительностью 30 секунд.

4.2 Описание эксперимента

Для эксперимента была выделена случайная подвыборка из 22 участников с равным количеством мужчин и женщин, и аудиофрагменты, которые они слушали, а также соответствующие записи ЭЭГ-сигналов. Все данные были разделены в соотношении 80:20. Объединение частей с начала сигнала было использовано в качестве обучающей выборки, а объединение частей с конца было использовано в качестве тестовой выборки (см. 3 и [1]).

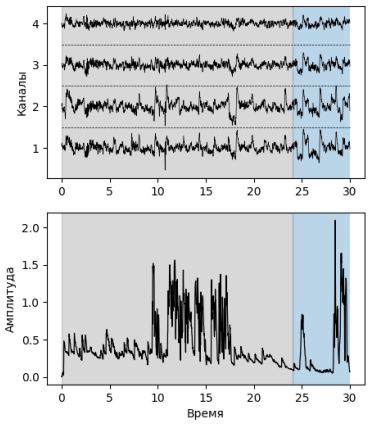

Рис. 3: Разделение на обучающую и тестовую выборки

Так как исходные обработанные ЭЭГ-сигналы и аудиофрагменты слишком велики, они были разбиты на окна фиксированной длины. Для каждой пары окон (ЭЭГ-сигнал, истинный стимул) были добавлены $K-1$ ложных стимулов из других пар. Далее для обеспечения сбалансированности классов были созданы $K-1$ копий каждого объекта со смещением истинного стимула.

С учетом языка аудиосигналов было решено взять предобученную модель wav2vec2-base-960h-phoneme-reco-dutch и whisper-small. Для каждого аудио с помощью модели создавался эмбеддинг, который потом выравнивался, по уже готовым данным.

В эксперименте ширина окна была взята равной 5 секунд с шагом 1 секунда. Количество стимулов $K=5$, то есть один истинный стимул и четыре ложных. Эксперимент проводился в 10 эпох, а размер батча составлял 64 элементов. В качестве метрики взято среднее по участникам долей правильных ответов.

Обозначим множество классов, как $\{0,\ldots,K-1\}$. Учитывая это, метрика качества вычисляется по формуле

$$Score = \frac{1}{22}\sum_{i=1}^{22}\frac{1}{l_i}\sum_{j=1}^{l_i}\left[y_j^i = \text{pred}_j^i\right],$$

где $y_j^i \in \{0,\ldots,K-1\}$ — метка объекта, l_i — количество кортежей для i-го участника, а pred_j^i — предсказание модели на объекте j.

4.3 Результаты эксперимента

Лучший результат был достигнут при комбинировании трансформера-кодировщика с Wav2Vec2. Также повысилось качество и при использовании Whisper-small. Результаты моделей представлены в таблице 1.

Model	Score (%)
Baseline	47.68 ± 11.75
Transformer Encoder	48.15 ± 10.33
Wav2Vec2	47.92 ± 11.54
Whisper-small	48.04 ± 9.85
Transformer Encoder + Wav2Vec2	**48.70** ± 9.44
Transformer Encoder + Whisper-small	48.36 ± 9.24

Таблица 1: Оценка качества на использованных моделях.

Диаграмма размаха качества классификации для каждого участника представлена на рис. 4. Заметим, что внедрение трансформера-кодировщике уменьшает количество выбросов и, в целом, повышает качество.

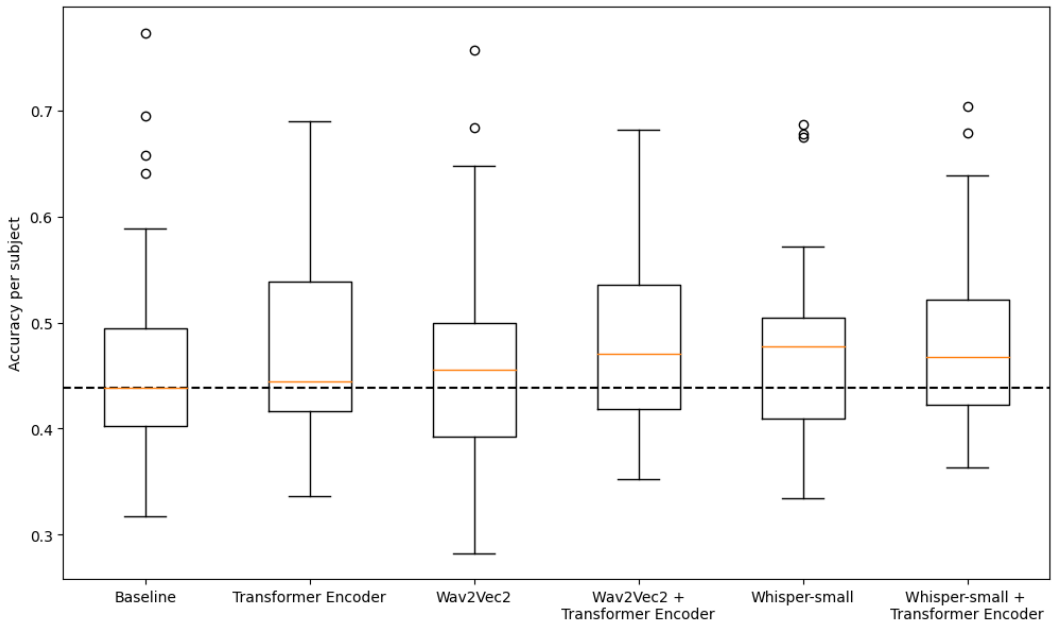

Рис. 4: Диаграмма размаха построенная на предсказаниях тестовых данных

Также стоит отметить, что межквартильный размах расположен выше базового решения у комбинированных моделей и трансформера-кодировщика. Наибольший прирост качества, который был получен благодаря трансформеру-кодировщику и модели Wav2Vec2, объясняется тем, что трансформер-кодировщик использует механизм внимания, и тем, что предобученная модель Wav2Vec2 была настроена на соответствующий язык аудиоданных. Использование эмбеддингов модели Whisper-small чуть уступает по качеству, так как предобученная модель является общей для всех языков и не настраивалась именно на нидерландский.

5 Заключение

Проведенные эксперименты продемонстрировали, что предложенные улучшения в архитектуре базовой модели позволили достичь повышения качества классификации стимулов для заданного ЭЭГ-сигнала. Внедрение трансформера-кодировщика вместо сверточной нейронной сети оказалось полезным для

улавливания зависимостей в ЭЭГ-сигналах, что способствовало более точному представлению данных в латентном пространстве и уменьшению количества выбросов.

Использование современных моделей автоматического распознавания речи, таких как Wav2Vec2 и Whisper, в качестве энкодера стимула также увеличило точность классификации. Наилучший результат был достигнут при использовании трансформера-кодировщика и модели Wav2Vec2.

На основе полученных результатов появляется вопрос об зависимости качестве модели от ширины окна и количества стимулов. Также предстоит провести анализ зависимости качества от размера скрытого состояния физико-информированных энкодеров, так как большая размерность позволяет хранить больше информации о речи. Дополнительно необходимо сравнить полученные результаты с другими физико-информированными энкодерами.

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] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020.

[4] 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.

[5] 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.

[6] Steffen Dasenbrock, Sarah Blum, Paul Maanen, Stefan Debener, Volker Hohmann, and Hendrik Kayser. Synchronization of ear-eeg and audio streams in a portable research hearing device. Frontiers in Neuroscience, 16, 09 2022.

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

[8] 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.

[9] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

[10] 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.

[11] 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.

[12] 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.

[13] Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth. Self-supervised models of audio effectively explain human cortical responses to speech, 2022.

[14] 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.

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

[16] 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.

[17] 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.