Использование методов глубокого обучения для обработки текстовых данных УЗИ щитовидной железы

Е.В. Боброва, О.И. Моисеенко, Е.В. Дюльдин, К.С. Зайцев, А.А. Гармаш, А.А. Трухин, С.М. Захарова, Е.А. Трошина

Аннотация. **Целью настоящей статьи является исследование различных интеллектуальных алгоритмов для обработки русскоязычной текстовой медицинской информации, полученной в результате УЗИ щитовидной железы, решении задач классификации заболеваний по системе EU-TIRADS и генерации заключения врача по описанию заболевания. В рамках исследования разработан конвейер машинного обучения, включающий этапы предобработки данных и обучения моделей. Для проектирования моделей глубокого обучения использовались трансформерные и гибридные архитектуры. В работе предложены методы предобработки неструктурированных медицинских описаний для их адаптации к требуемому формату решаемых задач. Полученные в ходе исследования результаты показали, что при решении задачи классификации достижение стабильных и высоких результатов с использованием нейросетевых архитектур возможно только при тщательном подборе гиперпараметров и учета их взаимного влияния. При решении задачи генерации заключений врача УЗИ трансформерные архитектуры и большие языковые модели показывают хорошие результаты на больших объемах данных. Предложенное решение в рамках программного комплекса «Интеллектуальный ассистент врача УЗИ» позволит автоматизировать труд врача и улучшить качество диагностики.**

Ключевые слова **– глубокое обучение, языковая модель, трансформер, щитовидная железа, УЗИ, классификация EU-TIRADS**

Статья получена 2024.

Боброва Елизавета Витальевна, Национальный Исследовательский Ядерный Университет МИФИ, магистрант, EVBobrova@mephi.ru

Моисеенко Олеся Игоревна, Национальный Исследовательский Ядерный Университет МИФИ, бакалавр, Olesya.moiseenko20@gmail.com

Дюльдин Евгений Владимирович, Национальный Исследовательский Ядерный Университет МИФИ, аспирант, zhecos1@yandex.ru

Зайцев Константин Сергеевич, Национальный Исследовательский Ядерный Университет МИФИ, профессор, KSZaytsev@mephi.ru

Гармаш Александр Александрович, Национальный Исследовательский Ядерный Университет МИФИ, директор Института биомедицины и физики, AAGarmash@mephi.ru

Трухин Алексей Андреевич, ФГБУ «НМИЦ эндокринологии» Минздрава России, медицинский физик, alexey.trukhin12@gmail.com

Захарова Светлана Михайловна, ФГБУ «НМИЦ эндокринологии» Минздрава России, врач ультразвуковой диагностики, smzakharova@mail.ru

Трошина Екатерина Анатольевна, ФГБУ «НМИЦ эндокринологии» Минздрава России, чл. корр., директор Института клинической эндокринологии, troshina@inbox.ru

I. ВВЕДЕНИЕ

УЗИ щитовидной железы является неинвазивной диагностической процедурой, которая позволяет оценить структуру и геометрию этой железы.

Показаниями для УЗИ щитовидной железы являются пальпируемые узлы в области шеи; изменения в размерах щитовидной железы; боль или дискомфорт в области шеи; контроль за уже известными патологиями щитовидной железы.

Для более точного определения характера обнаруженных при УЗИ узлов щитовидной железы и выбора дальнейшей тактики лечения, в Российской Федерации используют система классификации (категоризации) состояния железы European Thyroid Imaging Reporting and Data System (EU-TIRADS) [1]. Также существуют системы ACR-TIRADS, K-TIRADS, ATA рекомендации и др.

EU-TIRADS помогает стандартизировать интерпретацию данных УЗИ в разных странах и определить дальнейшую траекторию ведения пациента. Система включает оценку геометрических и структурных характеристики узлов щитовидной железы. В EU-TIRADS выделено пять категорий, коррелирующих с риском злокачественности образования. Применение системы в части оценки необходимости дальнейшего наблюдения, проведения пункционной биопсии активно исследуется экспертным сообществом.

В последние годы компьютерные методы глубокого обучения стали постепенно применяться и в эндокринологии для сокращения сроков и повышения точности диагностических процедур. Классификация, базирующаяся на глубоких нейронных сетях, позволяет существенно улучшить интерпретацию изображений и кино-петель УЗИ, проверить текстовое описание, созданное врачом УЗИ, и автоматически сгенерировать заключение.

Клиническое мышление можно представить как алгоритм последовательного представления объекта исследования в виде модели характеристик, сравнения установленных характеристик с

известными критериями и естественно-языковую интерпретацию.

Перед врачом-специалистом сегодня ставится задача успеть за прогрессом в области информационных технологий, подтвердить их эффективность и обеспечить преемственность ассоциированных медицинских знаний.

Применение таких методов в рамках системы EU-TIRADS может не только ускорить процесс диагностики, но и повысить его точность, минимизируя вероятность пропуска злокачественных узлов, снижая количество ненужных биопсий.

Целью данной работы является использование современных методов искусственного интеллекта (ИИ) для обработки естественно-языковых текстовых описаний, созданных исключительно врачом-человеком или с использованием им компьютерного ассистента [2] для классификации и генерации медицинских заключений в соответствии с системой EU-TIRADS. В ходе исследования будут рассмотрены различные решения этих двух задач, базирующиеся на современных архитектурах нейронных сетей и проанализирована их эффективность.

II. ИСХОДНЫЕ ДАННЫЕ

Исходные данные были собраны в течение семи лет врачами ФГБУ «НМИЦ эндокринологии» Минздрава России, и представляют собой обширный анонимизированный датасет, содержащий 7286 записей результатов ультразвукового исследования (УЗИ) щитовидной железы. Эти данные включают следующие основные категории: объемные образования: описание обнаруженных объемных образований в щитовидной железе; дополнительные данные: подробная информация, связанная с УЗИ, включая детали о расположении, размерах и других характеристиках обнаруженных образований; заключение: итоговое заключение по результатам УЗИ.

Пропущенные значения присутствуют в трех столбцах, особенно в "Дополнительных данных", где зафиксировано 5275 пропущенных значений. Анализ длины текстов показал, что средняя длина текста составляет 12,68 символа, с максимумом в 218 символов, что указывает на разнообразие объема информации в текстах.

Проведенный анализ топ биграмм и триграмм выявил ключевые темы в текстах, такие как "European Thyroid Imaging" и "Reporting and Data", что позволяет идентифицировать основные области фокусировки данных [3]. Для визуализации текстовых данных использовалось облако слов, которое эффективно демонстрирует ключевые слова и темы в документах (рис. 1). Это графическое представление отражает не только частоту использования слов, но и их взаимосвязи, формируя тематические кластеры и выявляя основные темы и подсистемы в тексте.

Облако слов позволяет анализировать текстовые данные в двух основных направлениях – распределение слов и взаимосвязи между словами [4].

При распределении слов учитываются частота использования и специфичность. Частота использования сообщает – слова, написанные более крупными буквами, в облаке слов указывают на наиболее часто встречающиеся термины в тексте, что помогает определить ключевые темы. Специфичность подсказывает, что менее распространенные слова могут быть более специфичными для контекста, раскрывая уникальные аспекты темы.

При анализе взаимосвязей между словами выделяют семантические связи и тематические кластеры. Семантические связи показывают, что слова, расположенные близко друг к другу, часто используются вместе, что демонстрирует их семантическую связь, и помогает выявлять ключевые идеи. Группировка слов вокруг центральных терминов создает тематические кластеры, отражающие различные аспекты обсуждаемых тем.

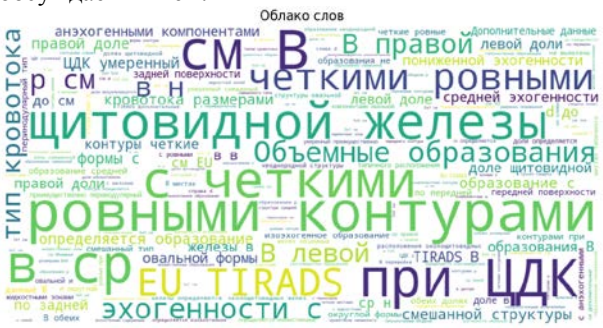

Рис. 1. Облако слов для данных УЗИ

Контекст и тематику данных можно оценить путем анализа сентимента с использованием средства VADER [5], которое показало положительную эмоциональную окраску текстов (среднее значение сентимента 0,235), при этом в текстах присутствуют как сильно позитивные, так и негативные окрашенные элементы. Для выявления основных тем использовался метод LDA (Latent Dirichlet Allocation), который позволяет определить пять ключевых тематических областей, включая описание объемных образований, контуры и размеры, а также расположение относительно других структур.

Результирующий анализ подтверждает сложность и многомерность данных, полученных в ходе УЗИ щитовидной железы. Пропущенные значения в некоторых категориях требуют особого внимания при дальнейшей обработке и анализе. Статистический анализ и анализ сентимента позволяют оценить эмоциональную окраску и уровень детализации текстовых данных, что важно для понимания контекста и интерпретации результатов исследования.

Также был проведен анализ вхождения меток EU-TIRADS. В таблице 1 представлено распределение меток EU-TIRADS в исходном наборе данных:

Таблица 1. Распределение меток EU-TIRADS

Класс	Количество вхождений	Процент вхождения метки EU-TIRADS
1	10	0,1%
2	1920	30,0%
3	3118	48,9%
4	1036	16,0%
5	377	5,0%

Анализ таблицы выявил, что общее количество вхождений меток в данные превышает количество записей. Это явление обусловлено присутствием записей без меток в столбце "Заключение", например, таких как "эхографические признаки многоузлового зоба". Кроме того, в одном заключении могут присутствовать несколько меток EU-TIRADS, как в приведенном примере: «Эхографические признаки двустороннего многоузлового зоба (EU-TIRADS 2 и 3 с обеих сторон и в перешейке), с частично загрудинным расположением справа». Также необходимо учитывать наличие неинформативных заключений, где часть информации хранится в столбцах "Объемные данные" или "Дополнительные данные".

III. ПРЕДОБРАБОТКА НАБОРА ДАННЫХ

Для успешного выполнения задач генерации и классификации, критически важна предварительная обработка данных. Эта процедура включала в себя следующие этапы

1. Очистка данных: удаление записей без меток и обработка неинформативных заключений, чтобы обеспечить единообразие и полноту данных. Включает в себя удаление или корректировку записей с отсутствующими или противоречивыми данными, что помогает избежать ошибок и искажений в дальнейшем анализе.

2. Нормализация данных: приведение различных форм представления меток к единой форме. Необходимо для того, чтобы все метки имели стандартный вид, что предотвращает ошибки при их интерпретации и анализе. Например, метки EU-TIRADS могут быть представлены в различных форматах, и их нормализация поможет унифицировать данные.

3. Разделение данных: разделение данных на обучающую, валидационную и тестовую выборки. Позволяет эффективно обучить модель и проверить ее на независимых данных. Стандартное соотношение составляет 70% для обучения, 15% для валидации и 15% для тестирования, что обеспечивает сбалансированное распределение данных для каждой из задач.

4. Кодирование категориальных переменных: преобразование категориальных переменных в числовые значения. Необходимо для того, чтобы алгоритмы машинного обучения могли обрабатывать категориальные данные. Например, метки EU-TIRADS могут быть закодированы в числовую форму, что облегчает их обработку моделями машинного обучения.

5. Обработка пропущенных значений: замена или удаление пропущенных значений. Улучшает качество данных и стабильность модели. Методы замены включают использование средних значений, медиан или предсказанных значений на основе других данных [6]. Также возможно удаление записей с пропущенными значениями, если их количество незначительно.

Медицинские тексты, как правило, обладают слабой структурой, что затрудняет удаление шума и выделение ключевой информации, необходимой для дальнейшего анализа и генерации признаков. Для решения задачи выделения целевых меток и ключевых токенов описания на неразмеченном множестве данных, предпочтительным методом является использование регулярных выражений и вероятностного поиска наиболее часто встречающихся токенов предложений. Регулярные выражения позволяют обнаруживать различные комбинации меток EU-TIRADS в текстах.

После очистки данных и формирования меток, предложения можно сгруппировать по классам. Этот процесс включает также определение количества предложений в тексте и вероятности появления слов в каждом предложении, что является важным параметром при различных методах аугментации данных и валидации результатов. Разделение длинных заключений на отдельные предложения с метками EU-TIRADS позволяет расширить признаковое поле на 8,4%, что значительно улучшает качество и точность модели.

Таким образом, тщательная предобработка данных является фундаментом для успешного применения методов глубокого обучения в задачах многоклассовой классификации и генерации медицинских данных по системе EU-TIRADS. Правильная очистка, нормализация и кодирование данных, а также обработка пропущенных значений и структурирование текстов обеспечивают основу для построения надежной и эффективной модели, способной решать сложные задачи медицинской диагностики.

IV. РЕШЕНИЕ ЗАДАЧИ КЛАССИФИКАЦИИ

Классификация данных, особенно в медицинской области, требует особого подхода, учитывающего специфику и сложность текстовых данных [7]. В данной работе используется модель глубокого обучения, способная эффективно обрабатывать и классифицировать медицинские тексты на основе системы EU-TIRADS.

Первая модель, основанная на рекуррентном подходе, включает в себя несколько ключевых компонентов, каждый из которых играет важную роль в процессе обучения и построении прогноза (табл.2).

Входной слой (Input layer): принимает последовательности текстовых данных. Размер

входных данных определяется параметром max_len, который ограничивает максимальную длину последовательности, передаваемой в модель.

Embedding слой: преобразует входные текстовые данные в числовые представления. Этот слой создает матрицу векторов, где каждому уникальному слову соответствует вектор фиксированного размера (128 в данном случае). Это позволяет модели работать с числовыми данными, что упрощает процесс обучения.

Bidirectional LSTM слой: LSTM (Long Short-Term Memory) — это тип рекуррентной нейронной сети (RNN), способный запоминать долгосрочные зависимости в данных. Бидирекциональный LSTM позволяет модели обрабатывать входные данные в обоих направлениях времени, что улучшает способность модели извлекать информацию из последовательностей [8].

Dropout слой: применяется для предотвращения переобучения модели, случайным образом исключая некоторое количество нейронов на каждом шаге обучения.

Conv1D слой: выполняет свертку вдоль временной оси, позволяя модели выявлять локальные паттерны в данных.

MaxPooling1D слой: уменьшает размерность пространства, выбирая максимальные значения из окон размером 2, что помогает уменьшить количество параметров и ускорить обучение.

BatchNormalization слой: нормализует выходные данные каждого временного шага, что помогает ускорить обучение и улучшить стабильность модели.

Dense слой с активацией ReLU: преобразует выходные данные из предыдущих слоев в более высокоразмерное пространство.

Выходной слой Dense с активацией softmax: используется для классификации входных данных в одно из 30 возможных классов.

Модель компилируется с использованием оптимизатора RMSprop [9] и функции потерь categorical_crossentropy, которая подходит для многоклассовой классификации. Метрики оценки включают точность, полноту и F1-оценку, что помогает оценить производительность модели.

Эта архитектура представляет собой комплексный подход к обработке и классификации текстовых данных, сочетая в себе преимущества различных типов слоев и техник, таких как LSTM, Conv1D и Dropout, для достижения высокой точности и обобщающей способности модели.

Далее было решено использовать модель GRU (Gated Recurrent Units). Сеть начинается с Embedding слоя, который преобразует входные данные — последовательности индексов слов — в плотные векторы. Этот шаг позволяет модели работать с числовыми данными, что упрощает дальнейшую обработку и обучение.

Затем следует Bidirectional GRU слой, который применяет двунаправленный рекуррентный процесс к каждому шагу входной последовательности. Это позволяет модели учитывать контекст как вперед, так и назад от текущего шага, улучшая тем самым обработку временных зависимостей в данных.

После Bidirectional GRU слоя, вводится слой Dropout, который случайно отключает некоторые нейроны во время обучения. Это важный механизм для предотвращения переобучения и повышения обобщающей способности модели.

Затем следует еще один GRU слой, который дополнительно обрабатывает последовательности, полученные от предыдущих слоев. Этот слой не имеет регуляризации, что позволяет ему сосредоточиться на обработке данных [10].

Далее, Dense слой с активацией ReLU преобразует выходные данные предыдущих слоев в пространство более высокого уровня, где можно легко вычислить вероятности принадлежности к классам. Этот слой также включает L1 и L2 регуляризацию для предотвращения чрезмерной зависимости от отдельных весов [11].

И, наконец, Output слой с активацией softmax преобразует внутреннее состояние модели в вероятностное распределение над всеми возможными классами, что позволяет интерпретировать выводы модели как вероятности принадлежности к каждому из классов.

Таблица 2. Детальная архитектура GRU подхода

Слой	Тип слоя	Параметры	Комментарии
Embed-ding	Вложе-ние слов	input_dim=max_words, output_dim=embedding_dim	Преобразует индексы слов в плотные векторы
Bidirecti onal GRU	Глубо-кий рекур-рентный слой	units=gru_units, return_sequences=True, kernel_regularizer=l1_l2(l1=0.05, l2=0.03), recurrent_regularizer=l1_l2(l1=0.01, l2=0.03), bias_regularizer=l1_l2(l1=0.01, l2=0.01)	Обработка последовательностей с учетом контекста вперед и назад
Drop-out	Отклю-чение	rate=dropout_rate	Предотвращение переобучения путем случайного отключения нейронов
GRU	Глубо-кий рекур-рентный слой	units=gru_units	Дополнительная обработка последовательностей
Dense	Полно-связный слой	units=dense_units, activation='relu', kernel_regularizer=l1_l2(l1=0.05, l2=0.03)	Преобразование данных в пространство более высокого уровня
Output	Выход-ной слой	units=output_size, activation='softmax', kernel_regularizer=l1_l2(l1=0.05, l2=0.03)	Вычисление вероятностей принадлежности к классам

Так же важно было использовать и испытать более сложные архитектуры для задачи классификации [12]

Авторы статьи [13] показали, что модель RuBioRoBERTa демонстрировала хорошие результаты при классификации медицинских текстов, однако одной из основных проблем при

дообучении моделей BERT является нестабильность результатов. Несмотря на то, что дообучение позволяет достигать высоких показателей точности, результаты могут значительно варьироваться. Основными причинами этой проблемы являются "забывчивость" нейронной сети и ограниченный объем данных для дообучения. Эту проблему исследуют уже длительное время, и существуют различные методы её решения, которые требуют отдельного рассмотрения.

Параметр инициализации генератора случайных чисел (seed) варьировался в диапазоне [2; 2048], что привело к различным результатам (табл. 3):

Таблица 3. Подбор параметров seed

Seed	2	4	8	16	32	64	128	256	512	1024	2048
Accu-racy	89.92	87.87	87.48	87.87	89.92	90.32	89.92	85.92	89.05	85.14	85.14

Несмотря на достижение высокой точности, диапазон значений (85-90%) показывает значительную вариативность, что, вероятно, связано с несбалансированностью классов.

Параметр learning rate (LR) определяет размер шага на каждой итерации при движении к минимуму функции потерь. Изменение LR приводит как к увеличению, так и к уменьшению точности. В лучшем случае точность удалось увеличить на ~1%. Важно отметить, что LR тесно связан с другим параметром – batch size, который определяет количество примеров, используемых для вычисления частичного значения функции потерь. Теоретически, изменение одного параметра в k раз требует изменения другого в k раз для улучшения результата, но на практике это не всегда работает.

Результаты экспериментов с batch size приведены в табл. 4.

Таблица 4. Подбор параметра batch size

Batch size	4	8	16	32
Accuracy	90.32	88.92	89.92	90.81

Как видно из таблицы, максимальная точность была достигнута при большем batch size, что уменьшило время обучения. Но, пропорциональное увеличение LR не улучшило результат, что еще раз подтвердило зависимость гиперпараметров от конкретных условий обучения.

Для достижения стабильных и высоких результатов важно тщательно подбирать гиперпараметры и учитывать их взаимное влияние. Экспериментально полученные параметры обучения: epochs = 25, batch size = 16, LR = 3e-5.

V. РЕШЕНИЕ ЗАДАЧИ ГЕНЕРАЦИИ

Для следующего этапа обучения – генерации текста заключения использовался набор данных по УЗИ щитовидной железы, в виде таблицы со столбцами “input” и “target”. В первом столбце каждой строки находятся тексты описания УЗИ вида:

«В правой доле почти на всю долю анэхогенное образование с перегородками, неоднородной структуры, со взвесью, р 4,1х3,6х2,6см; в н/полюсе анэхогенная зона д-и 0,5см (EU-TIRADS 2).Трахея смещена влево.»

Во втором, целевом столбце каждой строки находятся тексты диагностических заключений врача, вида:

«Эхографические признаки многоузлового зоба.»

Общее количество таких текстов - 7862. Для тонкой настройки моделей и последующего снятия с них метрик оценки качества обучения данные предварительно были разделены на обучающую и тестовую выборки с помощью функции train_test_split(), представленной в библиотеке scikit-learn [14]. Вся выборка была поделена в примерном соотношении 90/10, так, что данных для обучения было 6862, а для тестирования – 1000. Далее тренировочная выборка была преобразована в формат словаря, так как именно такой вид данных требует модель для настройки.

Следующим этапом после формирования выборок была подготовка модели. Подразумевается импортирование предварительно обученной большой языковой модели из библиотеки transformers от Hugging Face [15]. Hugging Face – компания, разрабатывающая инструменты для создания приложений с использованием машинного обучения.

Библиотека transformers предоставляет простой интерфейс для использования предобученных моделей, а также инструменты для их обучения или тонкой настройки. Она включает в себя модели для различных задач, в том числе для обработки естественного языка.

Однако, как было описано в предыдущем разделе, Large Language Models – «тяжелые» и ресурсоемкие модели для использования, поэтому применяется процедура квантизации. Квантизация сокращает затраты памяти и вычислений за счет представления весов и активаций с типами данных более низкой точности, такими как 8-битные или 4-битные целые числа. Это позволяет загружать крупные модели, которые обычно не помещаются в память, и ускорять вывод. Библиотека transformers поддерживает различные алгоритмы квантизации, в том числе BitsAndBytesConfig. Одной из ключевых особенностей этого алгоритма является возможность загрузки моделей в 4-битном квантовании [16]. Это можно сделать, установив аргумент load_in_4bit=True при вызове метода .from_pretrained(). Таким образом возможно сократить использование памяти примерно в четыре раза.

Далее необходимо импортировать нужную модель, используя описанный раннее метод квантизации. Чтобы использовать нужную архитектуру, можно обратиться к одному из «Автоклассов» от Hugging Face. Эти классы нужны, чтобы автоматически получать соответствующую модель по имени или

пути к предварительно обученным весам, конфигурации или словарю.

В настоящем исследовании использовался класс AutoModelForCausalLM, экземпляр которого будет создан как один из классов моделей библиотеки “transformers” с помощью метода класса from_pretrained() с указанием пути к модели на Hugging Face, экземпляра алгоритма квантизации и способа хранения загруженной модели.

Далее был загружен токенизатор, для этого применялся класс AutoTokenizer. С помощью метода класса from_pretrained() с указанием пути к модели для обучения, и был создан экземпляр класса AutoTokenizer.

Следующим этапом является определение параметров для обучения. Для этого использовался класс TrainingArguments с аргументами выходного каталога, размеров батчей для каждого ядра ускорителя для обучения и числа шагов обновления, для которых нужно накопить градиенты, и др.

После задания необходимых параметров обучения, был определен класс Trainer, который предоставляет API для полнофункционального обучения с помощью фреймворка PyTorch. Но, так как в данной работе требовалось точно настроить большую языковую модель, то использовался класс SFTrainer из библиотеки trl, который является оболочкой Trainer, и специально оптимизирован для работы с LLM, Он поддерживает LoRA и квантизацию для эффективного масштабирования до любого размера модели. SFTrainer содержит базовый цикл обучения, который поддерживает вышеуказанные функции.

Для создания экземпляра класса SFTrainer также необходимо задать LoraConfig – класса математического метода LoRA, который ускоряет точную настройку больших моделей, позволяя потреблять при этом меньше памяти. Во втором разделе подробно описан этот метод. Чтобы сделать точную настройку более эффективной в подходе LoRA нужно представить обновления веса с помощью двух меньших матриц (матриц обновлений) посредством разложения низкого ранга. Эти новые матрицы можно обучить адаптироваться к новым данным, сохраняя при этом общее количество изменений на низком уровне. Исходная весовая матрица остается замороженной и не подвергается дальнейшим корректировкам. Для получения окончательных результатов объединяются исходные и адаптированные веса.

После того, как все необходимые классы для тонкой настройки модели заданы, можно перейти к процессу тонкой настройки (тюнингу) с помощью метода train() экземпляра класса SFTrainer.

Первой обученной моделью стала Llama 3 8B, для осуществления этого процесса применялись все описанные выше классы, частично импортированные с помощью библиотеки unsloth [17, отличительной особенностью которой является то, что все ядра написаны на языке OpenAI Triton. Важно уточнить, что Triton – это язык программирования графических процессоров с открытым исходным кодом для нейронных сетей. Он позволяет достичь максимальной производительности оборудования с относительно небольшими усилиями, например, его можно использовать для написания ядер матричного умножения FP16, которые за счет этого будут в два раза более эффективными, чем эквивалентные реализации Torch.

Для дообучения использовалось множество гиперпараметров в экземплярах классов BitsAndBytesConfig, LoraConfig и SFTrainer, так как эта модель имеет большое количество параметров и требует немалого количества ресурсов для обучения. После тщательной подготовки модели к настройке, была произведена ее тонкая настройка.

График поведения ее функции ошибки на тренировочной выборке приведен на рис. 2.

Из графика видно, что обучение шло стабильно с постепенным уменьшением функции ошибки без резких ее скачков, наименьшее значение составило 0.328. Обучение происходило меньше, чем за одну эпоху – на 1000 шагах, во избежание переобучения, для наглядности эпоха разбита на несколько десятков тысяч шагов.

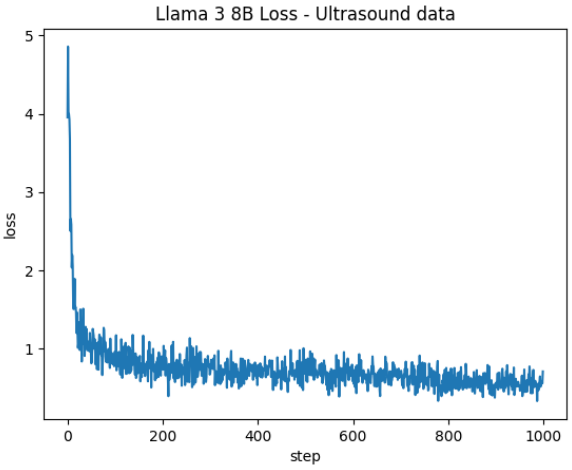

Рис. 2. График функции ошибки Llama 3 8B на данных УЗИ ЩЖ

Следующей тонко настроенной моделью была Mistral 7B. Дообучение происходило с применением всех описанных выше классов и методов для тонкой настройки LLM.

Tuning происходил в течение одной эпохи, разбитой на более чем 1000 шагов. Из графика на рис. 3 видно, что изменение функции потерь на тренировочной выборке относительно шагов обучения происходило в направлении приближения к минимуму, колебания значений происходили в пределах 0.3 значения функции потерь, причем минимальное значение составило 0.331. Можно сказать, что обучение проходило без атипичного поведения loss-функции. По поведению графика складывается ощущение, что разброс функции потерь очень сильный, однако так кажется лишь из-за маленького значения функции потерь (см. шкалу ординат) на первых шагах обучения.

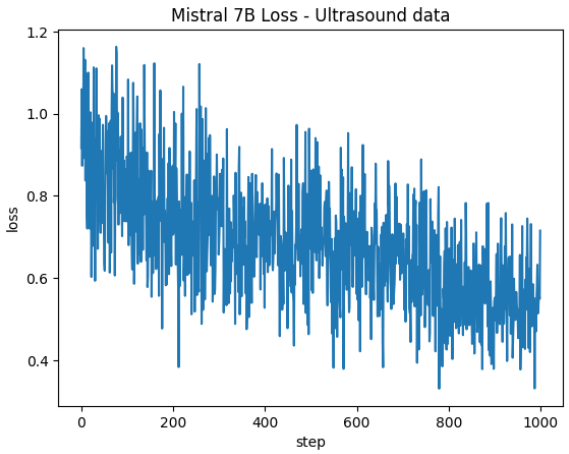

Рис. 3. График функции потерь Mistral 7B на данных УЗИ ЩЖ

VI. ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ И СРАВНЕНИЕ ИХ С РЕЗУЛЬТАТАМИ ГЕНЕРАЦИИ ТЕКСТА ПО СИСТЕМЕ BETHESDA

Проведенная работа по классификации и генерации текстовых документов цитологических [13] и ультразвуковых исследований щитовидной железы позволяет оценить производительность различных моделей машинного обучения в задачах этих нозологий. В таблице 5 представлены ключевые метрики качества моделей: Accuracy, Loss, Precision, Recall и F1-оценка, которые являются стандартными показателями эффективности классификационных моделей.

Проведем оценку эффективности работы проверенных нами моделей классификации заболеваний по шкалам EU-TIRADS (УЗИ) и Bethesda (цитологические исследования) при анализе текстовых описаний заболеваний щитовидной железы.

1. Гибридная модель (LSTM и Conv1D)

Для цитологических исследований гибридная модель показывает высокую Accuracy (0.715) и умеренные Loss (0.565). Однако, несмотря на высокие значения F1-оценки (0.525), показатель Precision (0.555) и Recall (0.505) указывают на возможные проблемы с балансом между ложноположительными и ложноотрицательными ошибками.

Для данных УЗИ гибридная модель демонстрирует немного меньшую Accuracy (0.655) и большую Loss (0.605) по сравнению с данными цитологии. Это может указывать на более сложные особенности данных УЗИ, требующие более тонкой настройки модели.

2. Модель GRU

Для данных цитологии модель GRU показывает высокую Accuracy (0.865) и низкие Loss (0.41), что свидетельствует о её высокой эффективности. Тем не менее, несмотря на высокую Precision (0.72) и Recall (0.665), возможны проблемы с балансом между типами ошибок.

Для данных УЗИ модель GRU показывает средние показатели Accuracy (0.725) по сравнению с цитологией, но лучше, чем у гибридной модели. Это говорит о том, что GRU лучше справляется с некоторыми особенностями данных УЗИ.

3. Модель BERT

Модель BERT для данных УЗИ демонстрирует исключительно высокие результаты в Accuracy (0.86) и Loss (0.225), а высокие значения Precision (0.88) и Recall (0.85) подтверждают эффективность использования контекстной информации для улучшения классификации.

Для данных УЗИ модель BERT показывает тоже достаточно высокие результаты Accuracy (0.755), хотя и ниже, чем для цитологии. Это может быть связано с более сложными текстовыми особенностями переданных нам данных УЗИ, и меньшим их общим количеством.

В целом, анализ показывает, что модели BERT и GRU дают лучшие результаты среди представленных моделей, особенно для цитологии (табл. 5). Гибридная модель показывает приемлемые результаты, но имеет потенциальные проблемы с балансом между типами ошибок.

Таблица 5. Сравнение производительности моделей

Модель / Тип данных	Accuracy	Loss	Preci -sion	Recall	F1-Score
Гибридная (Цитология)	0.715	0.565	0.555	0.505	0.525
Гибридная (УЗИ)	0.655	0.605	0.525	0.465	0.495
GRU (Цитология)	0.865	0.41	0.72	0.665	0.68
GRU (УЗИ)	0.725	0.495	0.635	0.595	0.615
BERT (Цитология)	0.86	0.225	0.88	0.85	0.86
BERT (УЗИ)	0.755	0.44	0.74	0.69	0.71

Таким образом, результаты показывают, что для решения задач классификации текстовых данных описаний заболеваний щитовидной железы в ультразвуковых и цитологических исследованиях лучше пользоваться моделями BERT и GRU. Гибридная модель может быть улучшена для более точного баланса между различными типами ошибок.

На графике, изображенном на рис. 4 представлено сравнение потерь (loss) модели Llama 3 при обучении на двух типах текстовых данных цитологических и ультразвуковых исследований. По оси абсцисс задано количество шагов обучения, по оси ординат - значения потерь. Из представленного графика можно сделать следующие выводы.

1. Наличие начальных потерь. В начале обучения оба типа данных демонстрируют высокие потери, причем данные ультразвуковых исследований (оранжевая линия) имеют более высокие начальные значения потерь, чем данные цитологических исследований (синяя линия).

2. Заметное снижение потерь. В течение первых 200 шагов обучения наблюдается резкое снижение потерь для обоих типов данных. Это

свидетельствует о быстрой адаптации модели к данным.

3. Стабилизация потерь: после начального быстрого снижения, потери стабилизируются. Однако, стабилизация происходит на разных уровнях: для данных цитологии потери стабилизируются на уровне около 0.5, тогда как для данных ультразвука - на уровне около 0.75.

4. Разница в эффективности: модель Llama 3 показывает более высокую эффективность при обработке данных цитологии по сравнению с данными ультразвука. Это выражается в более низких значениях потерь для данных цитологии на протяжении всего процесса обучения.

5. Шум в данных: данные ультразвука показывают большую вариативность и шум по сравнению с данными цитологии. Это может указывать на более высокую сложность текстовых данных ультразвука для обработки моделью.

В целом, модель Llama 3 демонстрирует лучшую производительность на данных цитологии по сравнению с данными ультразвука. Это подтверждается более низкими значениями потерь и меньшей вариативностью результатов для данных цитологии (рис. 4). Модель более эффективно справляется с задачами, основанными на данных цитологии, чем с задачами, основанными на данных ультразвука.

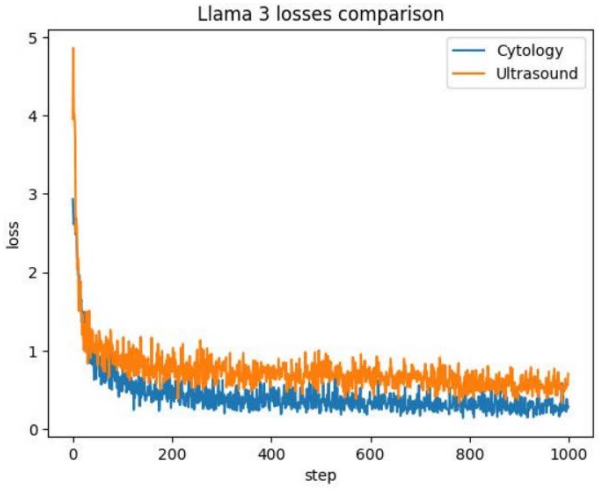

Рис. 4. График функции потерь Llama3

График на рисунке 5 представляет сравнение потерь (loss) другой модели - Mistral 7B при обучении на двух различных типах данных: цитологии и ультразвуке. Оси размечены, как в предыдущем случае. Выводы при анализе графика этого эксперимента аналогичны выводам по предыдущему графику.

1. Наличие начальных потерь. В самом начале обучения обе категории данных показывают высокие значения потерь. Однако, потери для данных ультразвука (оранжевая линия) изначально выше, чем для данных цитологии (синяя линия).

2. Заметное снижение потерь. С увеличением числа шагов обучения потери для обеих категорий данных снижаются, что свидетельствует об улучшении работы модели. По мере обучения модель учится лучше прогнозировать результаты, что отражается в снижении потерь.

3. Стабилизация потерь. По мере дальнейшего обучения потери стабилизируются. Однако, стабилизация происходит на разных уровнях: для данных цитологии потери стабилизируются на более низком уровне (примерно 0.25), тогда как для данных ультразвука они остаются выше (примерно 0.75).

4. Разница в эффективности. Наблюдаемая разница в уровнях потерь указывает на то, что модель Mistral 7B более эффективно обрабатывает данные цитологии по сравнению с данными ультразвука. Это можно объяснить более низкими значениями потерь и меньшей вариативностью в данных цитологии.

5. Шум в данных. Данные ультразвука демонстрируют большую вариативность и шум по сравнению с данными цитологии. Это может указывать на большую сложность или вариативность данных ультразвука, что затрудняет их обработку моделью.

В заключение, модель Mistral 7B показывает более высокую производительность на данных цитологии, что выражается в более низких потерях и меньшей вариативности результатов. Это свидетельствует о том, что модель более эффективно справляется с задачами, основанными на данных цитологии, по сравнению с задачами, основанными на данных ультразвука.

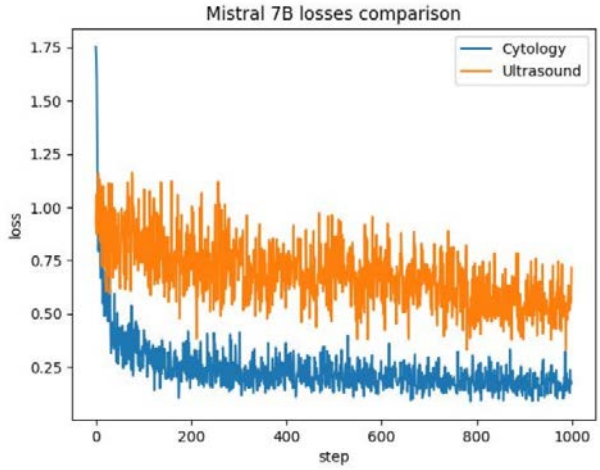

Рис. 5. График функции потерь Mistral 7B

Обе модели, Mistral 7B и Llama 3, показывают более высокую производительность на данных цитологии по сравнению с данными ультразвука. Однако модель Mistral 7B достигает более низких уровней потерь на данных цитологии, что свидетельствует о её лучшей общей производительности на этом типе данных. Модель Llama 3 демонстрирует более быстрое начальное снижение потерь, но её потери стабилизируются на более высоком уровне по сравнению с моделью Mistral 7B.

VII. ЗАКЛЮЧЕНИЕ

В результате проведенной работы можно констатировать.

При решении задачи классификации текстовых описаний ультразвуковых исследований заболеваний щитовидной железы достижение стабильных и высоких результатов возможно только при тщательном подборе гиперпараметров и учета их взаимного влияния. Экспериментально полученные параметры обучения следующие: число эпох - 25, размер батча - 16, Learning rate = 3e-5.

При решении задачи генерации текстового заключения по УЗИ обе используемые модели LLM (Mistral 7B и Llama 3) позволяют получить высокую точность при высокой производительности, при этом в экспериментах на реальных данных Mistral 7B все-таки несколько точнее. При обработке текстовых данных УЗИ используемые модели дают менее точный результат, измеренный по примененным метрикам, чем ранее при обработке данных цитологии [13], что связано с меньшим объемом доступного датасета УЗИ.

Представленные промежуточные результаты демонстрирует потребность в работе над лингвистическим обеспечением ультразвуковой диагностики узловых образований щитовидной железы, что позволит ускорить прогресс в области применения интеллектуальных алгоритмов в случае последовательного выполнения задач автоматизации лечебно-диагностических процессов.

Кроме того, повышение качества интеллектуальных алгоритмов может быть достигнуто методом машинного обучения с подкреплением, т.е. оценке специалистом сформированного компьютерной моделью заключения с последующей корректировкой алгоритма.

БЛАГОДАРНОСТИ

Авторы выражают благодарность Высшей инжиниринговой школе НИЯУ МИФИ за помощь в возможности опубликовать результаты выполненной работы и руководству ФГБУ «НМИЦ эндокринологии» Минздрава России за предоставленные текстовые данные.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Текстовые данные для проведения исследования подготовлены по гранту Российского научного фонда в рамках реализации проекта №22-15-00135 «Научное обоснование, разработка и внедрение новых технологий диагностики коморбидных йододефицитных и аутоиммунных заболеваний щитовидной железы с использованием возможностей искусственного интеллекта»

БИБЛИОГРАФИЯ

[1] Juhlin C. C., Baloch Z. W. The 3rd edition of Bethesda system for reporting thyroid cytopathology: Highlights and comments // Endocrine Pathology. – 2024. – Т. 35. – №. 1. – С. 77-79.

[2] Lozhkin I., Tsuguleva K., Zaytsev K. & oth. (2023) Development of Neural Network Models for Obtaining Information About Nodular Neoplasms of the Thyroid Gland Based on Ultrasound Images/ Journal of Theoretical and Applied Information Technology, 15th August 2023 - Vol. 101. No. 15—2023 p.p. 6076-6091.

[3] Egger, R., Gokce, E. (2022). Natural Language Processing (NLP): An Introduction. In: Egger, R. (eds) Applied Data Science in Tourism. Tourism on the Verge. Springer, Cham.

[4] Wahdan, A., Salloum, S.A., Shaalan, K. (2022). Qualitative Study in Natural Language Processing: Text Classification. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. ICETIS 2021. Lecture Notes in Networks and Systems, vol 322. Springer

[5] Wahdan, A., Salloum, S.A., Shaalan, K. (2022). Qualitative Study in Natural Language Processing: Text Classification. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. ICETIS 2021. Lecture Notes in Networks and Systems, vol 322. Springer

[6] Wang, Z., Ezukwoke, K., Hoayek, A. et al. Natural language processing (NLP) and association rules (AR)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. J Intell Manuf (2023).

[7] Parsaeimehr, E., Fartash, M. & Akbari Torkestani, J. Improving Feature Extraction Using a Hybrid of CNN and LSTM for Entity Identification. Neural Process Lett 55, 5979–5994 (2023)

[8] Prusty, S., Patnaik, S., Sahoo, G., Rautaray, J., Prusty, S.G.P. (2024). Unstructured Text Classification Using NLP and LSTM Algorithms. In: Nakamatsu, K., Patnaik, S., Kountchev, R. (eds) AI Technologies and Virtual Reality. AIVR 2023. Smart Innovation, Systems and Technologies, vol 382. Springer

[9] Zou F. et al. A sufficient condition for convergences of adam and rmsprop //Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. – 2019. – С. 11127-11135.

[10] Poluru, Eswaraiah & Syed, Hussain. (2023). A Hybrid Deep Learning GRU based Approach for Text Classification using Word Embedding. EAI Endorsed Transactions on Internet of Things. 10. 10.4108/eetiot.4590.

[11] Demir-Kavuk O. et al. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features //BMC bioinformatics. – 2011. – Т. 12. – С. 1-10

[12] Yvon, F. (2023). Transformers in Natural Language Processing. In: Chetouani, M., Dignum, V., Lukowicz, P., Sierra, C. (eds) Human-Centered Artificial Intelligence. ACAI 2021. Lecture Notes in Computer Science(), vol 13500. Springer

[13] Diuldin E., Makanov A., Shifman B., Bobrova E., Osnovin S., Zaytsev K., Garmash A., Abdulkhabirova F. (2024). Using deep learning to generate and classify thyroid cytopathology reports according to the Bethesda system. Revue d'Intelligence Artificielle, Vol. 38, No. 2, pp. 729-737. https://doi.org/10.18280/ria.380237

[14] Pedregosa F. et al. Scikit-learn: Machine learning in Python //the Journal of machine Learning research. – 2011. – Т. 12. – С. 2825-2830.

[15] Jain S. M. Hugging face //Introduction to transformers for NLP: With the hugging face library and models to solve problems. – Berkeley, CA : Apress, 2022. – С. 51-67.

[16] Fasoli A. et al. 4-bit quantization of LSTM-based speech recognition models //arXiv preprint arXiv:2108.12074. – 2021.

[17] Sepulveda E. J. B. et al. Towards Enhanced RAC Accessibility: Leveraging Datasets and LLMs //arXiv preprint arXiv:2405.08792. – 2024.

.

Using deep learning methods to process text data from thyroid ultrasound

E.V. Bobrova, O.I. Moiseenko, E.V. Dyuldin, K.S. Zaitsev,
A.A. Garmash, A.A. Trukhin, S.M. Zakharova, E.A. Troshina

Abstract. The purpose of this article is to study various intelligent approaches to processing Russian-language textual medical information obtained as a result of ultrasound of the thyroid gland, solving problems of classifying diseases according to the EU-TIRADS system and generating a doctor's conclusion based on the description of the disease. As part of the research, a machine learning pipeline was developed, including the stages of data preprocessing and model training. Transformer and hybrid architectures have been used to design deep learning models. The paper proposes methods for preprocessing unstructured medical descriptions to adapt them to the required format of the tasks being solved. The results obtained during the study showed that when solving a classification problem, achieving stable and high results using neural network architectures is possible only with careful selection of hyperparameters and taking into account their mutual influence. When solving the problem of generating ultrasound doctor's reports, transformer architectures and large language models show good results on large volumes of data. The proposed solution within the framework of the "Intelligent Ultrasound Physician Assistant" software package will automate the doctor's work and improve the quality of diagnosis.

Keywords – deep learning, language model, transformer, thyroid gland, ultrasound, EU-TIRADS classification

REFERENCES

[1] Juhlin C. C., Baloch Z. W. The 3rd edition of Bethesda system for reporting thyroid cytopathology: Highlights and comments //Endocrine Pathology. – 2024. – Т. 35. – №. 1. – С. 77-79.

[2] Lozhkin I., Tsuguleva K., Zaytsev K. & oth.(2023) Development of Neural Network Models for Obtaining Information About Nodular Neoplasms of the Thyroid Gland Based on Ultrasound Images/ Journal of Theoretical and Applied Information Technology, 15th August 2023 -- Vol. 101. No. 15, 2023 p.p. 6076-6091.

[3] Egger, R., Gokce, E. (2022). Natural Language Processing (NLP): An Introduction. In: Egger, R. (eds) Applied Data Science in Tourism. Tourism on the Verge. Springer, Cham.

[4] Wahdan, A., Salloum, S.A., Shaalan, K. (2022). Qualitative Study in Natural Language Processing: Text Classification. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. ICETIS 2021. Lecture Notes in Networks and Systems, vol 322. Springer

[5] Wahdan, A., Salloum, S.A., Shaalan, K. (2022). Qualitative Study in Natural Language Processing: Text Classification. In: Al-Emran, M., Al-Sharafi, M.A., Al-Kabi, M.N., Shaalan, K. (eds) Proceedings of International Conference on Emerging Technologies and Intelligent Systems. ICETIS 2021. Lecture Notes in Networks and Systems, vol 322. Springer

[6] Wang, Z., Ezukwoke, K., Hoayek, A. et al. Natural language processing (NLP) and association rules (AR)-based knowledge extraction for intelligent fault analysis: a case study in semiconductor industry. J Intell Manuf (2023).

[7] Parsaeimehr, E., Fartash, M. & Akbari Torkestani, J. Improving Feature Extraction Using a Hybrid of CNN and LSTM for Entity Identification. Neural Process Lett 55, 5979–5994 (2023)

[8] Prusty, S., Patnaik, S., Sahoo, G., Rautaray, J., Prusty, S.G.P. (2024). Unstructured Text Classification Using NLP and LSTM Algorithms. In: Nakamatsu, K., Patnaik, S., Kountchev, R. (eds) AI Technologies and Virtual Reality. AIVR 2023. Smart Innovation, Systems and Technologies, vol 382. Springer

[9] Zou F. et al. A sufficient condition for convergences of adam and rmsprop //Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. – 2019. – С. 11127-11135.

[10] Poluru, Eswaraiah & Syed, Hussain. (2023). A Hybrid Deep Learning GRU based Approach for Text Classification using Word Embedding. EAI Endorsed Transactions on Internet of Things. 10. 10.4108/eetiot.4590.

[11] Demir-Kavuk O. et al. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features //BMC bioinformatics. – 2011. – Т. 12. – С. 1-10

[12] Yvon, F. (2023). Transformers in Natural Language Processing. In: Chetouani, M., Dignum, V., Lukowicz, P., Sierra, C. (eds) Human-Centered Artificial Intelligence. ACAI 2021. Lecture Notes in Computer Science(), vol 13500. Springer

[13] Diuldin E., Makanov A., Shifman B., Bobrova E., Osnovin S., Zaytsev K., Garmash A., Abdulkhabirova F. (2024). Using deep learning to generate and classify thyroid cytopathology reports according to the Bethesda system. Revue d'Intelligence Artificielle, Vol. 38, No. 2, pp. 729-737. https://doi.org/10.18280/ria.380237

[14] Pedregosa F. et al. Scikit-learn: Machine learning in Python //the Journal of machine Learning research. – 2011. – Т. 12. – С. 2825-2830.

[15] Jain S. M. Hugging face //Introduction to transformers for NLP: With the hugging face library and models to solve problems. – Berkeley, CA : Apress, 2022. – С. 51-67.

[16] Fasoli A. et al. 4-bit quantization of LSTM-based speech recognition models //arXiv preprint arXiv:2108.12074. – 2021.

[17] Sepulveda E. J. B. et al. Towards Enhanced RAC Accessibility: Leveraging Datasets and LLMs //arXiv preprint arXiv:2405.08792. – 2024.