ФИЗМАТ, 2023, том 1, № 1, с. 65–73

УДК 004.891.3

ПРОТОТИП СИСТЕМЫ РАСПОЗНАВАНИЯ УСТАЛОСТИ ПО ВИДЕО-, АУДИО- И ТЕКСТОВЫМ ДАННЫМ

© 2023 г. Д. А. Вейценфельд[1,2], Г. А. Киселев[1,*], Я. С. Коровин[3], С. В. Маков[4]

[1]ФИЦ "Информатика и управление" Российской академии наук", Москва, Россия

[2]Российский Университет дружбы народов им. Патриса Лумумбы, Москва, Россия

[3]Научно-исследовательский институт многопроцессорных вычислительных и управляющих систем, г. Таганрог, Россия

[4]Институт сферы обслуживания и предпринимательства (филиал) ДГТУ в г. Шахты, Шахты, Россия

**e-mail: kiselev@isa.ru*

Поступила в редакцию 16.09.2023 г.
После доработки 14.11.2023 г.
Принята к публикации 14.11.2023 г.

Описан прототип системы, использующей видео-, аудио- и текстовые данные для распознавания состояния усталости и низкой работоспособности человека. Для этого также была изучена и подробно описана задача VQA, а также особенности ее реализации на примерах из других исследований. Проведены эксперименты на наборах с большой вариацией задач: стандартная задача VQA на наборе VQA v2, сложные сцены CLEVR CoGenT, анализ кассовых чеков Receipt-AVQA-2023.

DOI: 10.56304/S2949609823010045, **EDN:** JVEDRC

ВВЕДЕНИЕ

Проблема усталости и эмоционального выгорания сотрудников является актуальной на многих рабочих местах, таких как оператор, машинист, водитель,где от человека требуется действовать строго по инструкциям, а рабочие смены длительны и монотонны. На таких рабочих местах часто присутствует повышенная опасность как для сотрудников, так и для клиентов, а плохое эмоциональное состояние сотрудника может привести к трагедии.

Традиционным подходом к этой проблеме является обращение к психологу. Но исследования психологов субъективны, в большинстве случаев сотрудник готовится к взаимодействию с психологом заранее и может продемонстрировать не актуальные данные, испытывая угрозу увольнения или изменения условий оплаты труда в худшую сторону. Существуют исследования, позволяющие по уровню гормонов (кортизол и мелатонин) и состоянию аминокислот в сыворотке крови определить уровень выгорания как производной от усталости, стресса, нарушения сна, выработки инсулина и т. д.

С развитием информационных технологий становятся актуальны автоматические системы ранней диагностики признаков усталости и эмоционального выгорания. Но осуществить диагностику на основе снятия биоматериалов в большинстве случаев невозможно по той же причине, что и в случае с обращением к психологу – сотрудник может к этому заранее подготовиться и принудительно привести себя в порядок только лишь на момент диагностики. Поэтому мы разрабатываем систему, позволяющую по видео-, аудио- и текстовым данным исследовать работников в стратегически важных областях и выявлять наличие стресса у испытуемого с высокой долей вероятности. Наша система носит не диагностический, но рекомендательный характер, оставляя выбор влияния на состояние испытуемого за его руководством. Одним из подходов мультимодального анализа является использование механизма Visual Question Answering (далее – VQA).

Задача в рамках настоящего исследования – протестировать методы мультимодального анализа данных и представить набор архитектур, которые будут в дальнейшем использоваться при разработке прикладного алгоритма анализа психоэмоционального состояния человека. Рассмотрим подробнее упомянутый ранее мультимодальный механизм – VQA.

Впервые механизм VQA был представлен в [1] и развит в [2–13]. Существует много различных реализаций механизма, и все они выполнены посредством глубоких нейронных сетей. Среди решений с открытым исходным кодом [5, 10–12], наиболее подходящее для исследования и экспериментов над ним – Oscar [11], ввиду большой документированности и хороших результатов на соревновании VQA Challenge.

Для создания эффективного механизма мультимодального анализа пользователей необходимо учитывать фактор модульности системы и ее адаптации со сторонним обеспечением. Наша модель KVQA разрабатывается на основе Oscar, улучшая и модернизируя ее параметры. Основными техническими требованиями механизма являются возможности по анализу текстовой, аудио- и видеоинформации пользователя и выработка решения относительно его усталости.

Идентификация эмоциональной окраски произносимой речи без семантического анализа чаще всего осуществляется с помощью CNN, рекурентной сверточной искуственной нейронной сети (RNN) и LSTM с использованием end-to-end обучения. В качестве входных данных для решения такой задачи наиболее часто применяют векторы, представляющие собой MFCC с размерностью от 24 до 39. Также есть попытки применять вектора, получаемые при помощи DNN, на вход которой поступает непосредственно интервал оцифрованного речевого сигнала или векторы MFCC, а на выходе так называемые X-vectors, содержащие значимую для решаемой задачи информацию.

Так же, немаловажной частью анализа естественного языка, является выявление эмоций говорящего из речевой информации без семантической привязки. Произносимые слова могут быть эмоционально окрашены с помощью таких средств, как интонирование, изменение интенсивности и темпа речи человека.

1. ИСХОДНЫЕ УРАВНЕНИЯ

Задача VQA заключается в создании модели машинного обучения, которая способна отвечать на вопросы о содержании изображения. Входными данными для задачи VQA являются изображения и соответствующие им вопросы, заданные на естественном языке. Выходными данными является один или несколько ответов, также в форме текста на естественном языке.

Постановка задачи VQA может быть формализована следующим образом:

Даны: Изображение I, Вопрос Q на естественном языке.

Требуется: Ответ A на естественном языке, связанный контекстом с вопросом Q и изображением I.

Задача VQA является кросс-модальной, так как требует от модели сочетания обработки изображения и естественного языка. В отличие от традиционных задач компьютерного зрения, в которых модель должна выделить определенные объекты или характеристики на изображении, задача VQA требует понимания контекста и содержания изображения, а также соответствующих ему вопросов на естественном языке.

В подавляющем большинстве работ, модели VQA состоят из трех модулей:

1. Модуль компьютерного зрения.

2. Модуль обработки естественного языка.

3. Кросс-модальный модуль.

Первые два модуля – предобученные, третий – обучаемый.

Модуль компьютерного зрения использует сверточные нейронные сети (CNN, Convolutional Neural Network) для извлечения признаков из изображений. В частности, используются предобученные CNN, такие как VGG (Visual Geometry Group), ResNet, Inception и другие, которые были обучены на больших наборах данных, таких как ImageNet.

В большинстве случаев, для обработки естественного языка используется архитектура трансформеров [14] (transformer), реже – LSTM (Long Short-Term Memory). Обе архитектуры хорошо работают с последовательными данными, такими как текстовые данные. В большинстве случаев, в качестве этого модуля используют Bert [15].

Кросс-модальный модуль получает образы объектов и токены текста из первых двух модулей соответственно, а затем синтезирует ответ. Ответ чаще всего синтезируется глубокой нейронной сетью (DNN, Deep Neural Network) для многоклассовой классификации. Выходным классом является метка ответа на множестве ответов в обучающем наборе данных. Эта DNN является основным агентом в задаче машинного обучения. Для более узких задач VQA, например когда тре-

буется только числовой ответ, DNN для классификации можно заменить на другой тип – для интерполяции, а также и для задач бинарной классификации. Для задач с многоклассовой классификацией, можно оставить исходный тип DNN, но заменить метки ответов в обучающем наборе данных на метки требуемого набора классов.

Для решения различных задач архитектура KVQA предполагает модульность системы. Таким образом, можно заменять предобученные части модели на другие, более подходящие для узкоспециализированных задач. Также у KVQA есть API, который позволяет встроить ее в другую систему. В модели используются наиболее свежие версии подмоделей, представленные HuggingFace [16].

За обработку текста отвечает токенизатор Bert [15], который используется в большинстве моделей VQA и уже довольно долго является эталонной моделью обработки естественного языка.

Bert (Bidirectional Encoder Representations from Transformers) – это языковая модель на основе трансформерной архитектуры (Vaswani и др. [14]), разработанная командой исследователей Google в 2018 году. Bert обучается на больших объемах текстовых данных и может использоваться для решения различных задач обработки естественного языка, таких как классификация текстов, ответ на вопросы, анализ тональности, машинный перевод и т. д. Bert также является предварительно обученной моделью, что позволяет использовать ее для тонкой настройки на конкретную задачу и улучшения результатов.

Bert является комплексной моделью с множеством модулей и подмодулей. Из них, для задачи VQA нужны два модуля: токенизатор и кодировщик.

Токенизатор Bert использует подход WordPiece [17], который заключается в разбиении слов на N-граммы и использовании их в качестве токенов. Токенизация происходит на уровне символов, после чего символьные последовательности преобразуются в последовательности N-грамм. N-Граммы могут быть представлены как отдельные символы (например, буквы) или как комбинации символов (например, биграммы или триграммы).

Токенизатор использует словарь, который состоит из всех возможных N-грамм, полученных в процессе обучения модели на большой выборке текстов. Это позволяет модели обрабатывать неизвестные ей слова, разбивая их на известные подслова. В модели KVQA, токенизатор Bert является предобученным модулем.

Кодировщик Bert (далее – BertEncoder) является обучаемым подмодулем в модели KVQA. Состоит из множества параллельных слоев Transformer, каждый из которых выполняет многоканальную обработку входных токенов.

Каждый слой Transformer в BertEncoder состоит из двух подслоев: механизма внимания и полносвязанного прямого прохода. Механизм внимания позволяет модели обращать внимание на различные взаимодействия между токенами во входной последовательности и присваивать им различный вес. Полносвязный прямой проход принимает на вход взвешенную сумму результатов механизма внимания и преобразует ее в новое представление последовательности.

BertEncoder обучается с использованием метода обратного распространения ошибки, минимизируя заданную функцию потерь на задаче, например задаче предсказания следующего слова или задаче классификации текста. BertEncoder позволяет модели Bert достичь высоких результатов в задачах обработки естественного языка, таких как вопросно-ответная система (наш случай), классификация текста и других.

Для кодировки изображений в KVQA используется модель ResNeXt-152 C4 (далее – C4). Является предобученным подмодулем системы. Используется реализация из репозитория MMF [18], предобученная на композитном наборе данных, состоящем из MSCOCO [19], OpenImages V4 [20], Objects365 [21] и Visual Genome [22]. Данная модель использовалась в VinVL [12], логическом продолжении Oscar.

Основная идея ResNeXt-152 заключается в использовании блоков с наиболее эффективными комбинациями фильтров для изучения характеристик изображений на разных уровнях. Благодаря этому модель может достичь высокой точности при решении сложных задач компьютерного зрения.

ResNeXt-152 имеет 152 слоя и состоит из нескольких блоков, в которых применяются операции свертки, множественной (batch) нормализации и активации. Кроме того, модель использует стратегию обучения на основе глубокой свертки (deep supervision), которая позволяет получить более стабильную и быструю сходимость в процессе обучения. C4 – это вариант модели, который, в отличие от стандартной ResNet, имеет четыре последовательных сверточных слоя в каждом блоке.

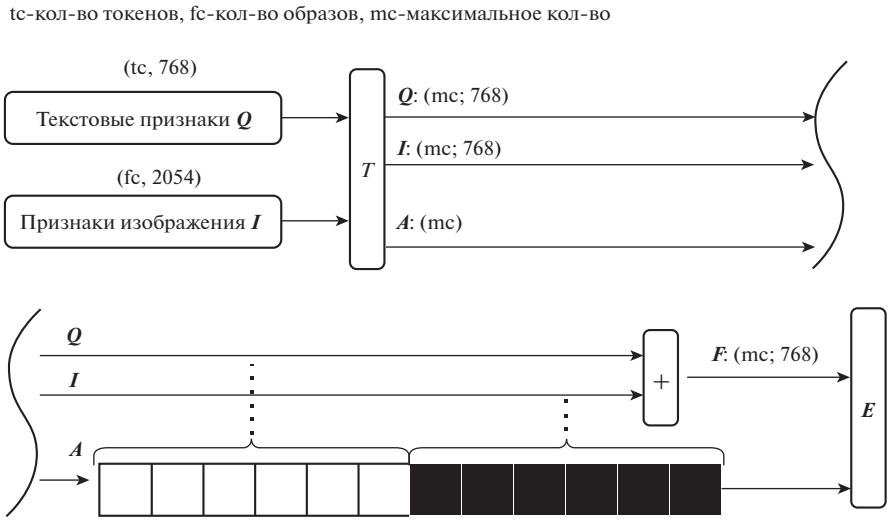

Рис. 1. Схема потока данных в кросс-модальном модуле KVQA..

Кросс-модальный модуль является самым важным для системы. Он представлен в виде двух моделей:

1. Модели преобразования T текстовых и визуальных признаков к матрицам одинаковой размерности, объединения их в одну.

2. Модели глубокой полносвязной нейронной сети для многоклассовой классификации.

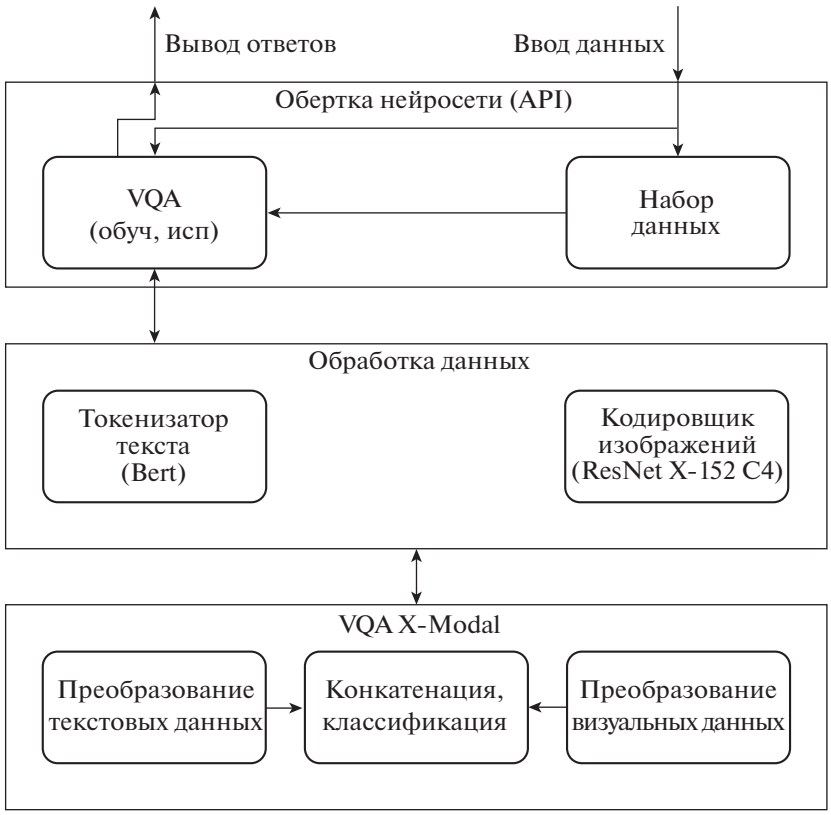

Рис. 2. Схема модели KVQA.

Таблица 1. Результаты тестирования модели KVQA

Набор	Точность
Oscar train dataset	72.8%
CLEVR CoGenT (A)	71.4%
CLEVR CoGenT (B)	65.7%
Receipt-AVQA-2023 (C4)	14.5%
Receipt-AVQA-2023 (CRAFT+Parseq)	23.2%

На модель преобразований приходят закодированные признаки текста вопроса Q и изображения I. Оба вектора признаков приводятся к одинаковой размерности. Затем признаки объединяются в один вектор, при этом генерируется маска внимания A (attention mask). Далее объединенные признаки F передаются на вход кодировщику BertEncoder E, который в свою очередь использует маску внимания, разделяющую визуальные и текстовые данные (рис. 1). Таким образом, BertEncoder, предназначенный изначально для кодировки текстовых данных, может оперировать точно так же и с данными изображения. Общая схема архитектуры KVQA изображена на рис. 2. Схема поделена на три блока, сверхувниз:

- Обертка модели, отвечающая за загрузку данных и использование их для обучения и оценки модели. Также в ней реализован сам алгоритм обучения.
- Предобученные модули модели.
- Обучаемые модули модели.

Один из тестовых наборов данных – AVQA-2023, требует анализ текста на изображении и числовой ответ. Набор данных AVQA представляет собой изображения кассовых чеков. Вопросы по кассовым чекам касаются задач подсчета товаров, суммы стоимости, а также поиска отдельной информации на чеке, такой как сумму налога. Все ответы являются числовыми с точностью до двух знаков после запятой.

Так как обычная модель для распознавания образов не подходит для чтения текста с изображения, для решения этой задачи было принято решение заменить стандартный для системы KVQA модуль зрения C4 на тандем из двух решений с открытым исходным кодом: CRAFT [23] и PArSeq [24].

Модель CRAFT (Character Region Awareness for Text detection) от ClovaAI является нейросетевой моделью для обнаружения текста на изображениях. Она используется для извлечения текстовых регионов из изображений и может быть применена в различных задачах, таких как распознавание текста, автоматическое преобразование рукописного текста в печатный и т. д. В KVQA эта модель используется для вырезания регионов с текстом для последующего распознавания в Parseq.

Модель Parseq от компании Baudm является нейросетевой моделью для выполнения естественно-языковых задач, таких как распознавание именованных сущностей, анализ зависимостей, заполнение пропущенных слов, генерация текста и т. д. В KVQA используется для распознавания текста с последующей кодировкой в основном модуле VQA.

Как было указано выше, задача VQA сводится к многоклассовой классификации, а значит мы можем заменить текстовые ответы на классификацию эмоционального состояния субъекта (сотрудника). Благодаря этому мы можем отдельно рассмотреть задачу распознавания эмоций по аудио, чтобы позже внедрить подходящую для этого модель машинного обучения в нашу систему KVQA. Для тестовой классификации эмоций по аудио была построена простая сверточная нейронная сеть из трех слоев свертки и одного полносвязанного слоя.

Методика анализа пользователя зависит от места работы. Если работник часто ведет диалог с диспетчером или клиентом, можно использовать изображение и звук прямо с рабочего места. Иначе, необходимо организовывать периодический (например, раз в месяц) опрос работника. Иной вариант – если позволяет устав на рабочем месте, и это указано в договоре об обработке персональных данных сотрудников – вести запись видео и звука с каких-либо зон отдыха, где сотрудники могут разговаривать между собой.

Во всех этих случаях текстовые данные должны присутствовать только с технической стороны и являться заготовленными вопросами о состоянии работника, и на основании ответов системы на них должен строиться вердикт об общем состоянии сотрудника. Формула (1) отражает опрос

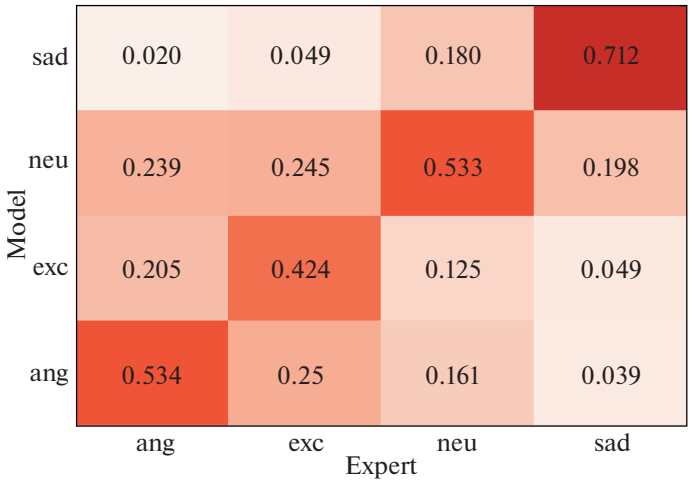

Рис. 3. Матрица ошибок определения эмоций с применением RNN.

модели KVQA. Здесь **Q** – набор вопросов, **I** – набор входных пар визуальных и аудиоданных (снятых с одного сотрудника в разные временные отрезки), **C** – набор пар вопрос-ответ, обозначающих ответы, к которым требуется внимание. Если условие (1) выполняется, то опрошенный сотрудник, вероятнее всего, имеет признаки эмоционального выгорания.

$$\forall \mathbf{Q}_i \in \mathbf{Q}, \quad \forall \mathbf{I}_i \in \mathbf{I} : ((\mathbf{Q}_i, Vqa(\mathbf{Q}_i, \mathbf{I}_i))$$

Теперь рассмотрим детальнее задачу анализа эмоций человека. В работе [25] для идентификации 4-х эмоциональных состояний говорящего (sad – грусть, neu – нейтральное состояние, exc – возбужденное состояние, ang – злость) используется рекуррентная нейронная сеть (RNN) с входными данными в виде 31-мерных MFCC-векторов. Полученные в работе результаты приведены на рис. 3.

Видно, что описанный метод определяет нейтральное состояние довольно часто для случаев возбужденного состояния и злости. Также довольно большое пересечение злости и возбужденного состояния.

Интересно, что даже автор описываемого метода довольно часто не совпал с мнением экспертов при выборе эмоции, соответствующей прослушиваемому фрагменту (рис. 4).

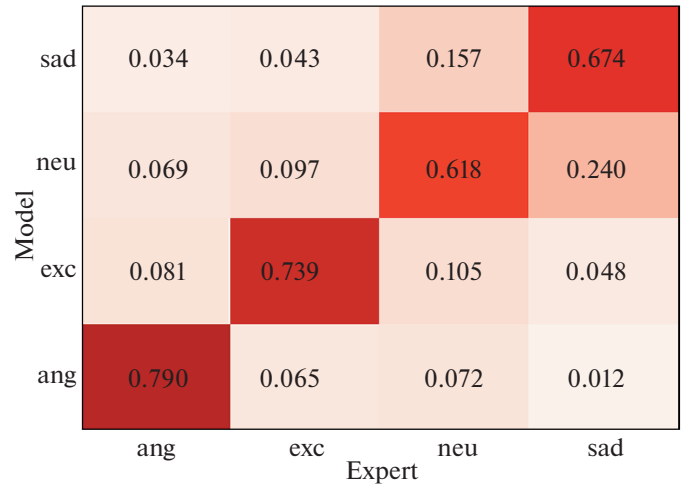

Рис. 4. Матрица ошибок определения эмоций самим автором метода.

Схожие результаты получены в работе [26]. В этой работе была сделана попытка выявления “токсичной” речи по анализу аудиосигнала. Полученная точность определения “токсичной” речи составила около 70%.

Целевая задача требует анализа состояния сотрудника на большом промежутке времени, а ее результаты должны проверяться лишь в редких периодах. Поэтому, от системы не требуется моментальной реакции в режиме real-time – требуется лишь периодически проверять сотрудника на предмет эмоционального выгорания. Исходя из этого, модели достаточно предоставлять результат анализа довольно редко – от нескольких минут до полного времени рабочей смены, благодаря чему аппаратные характеристики хост-машины модели могут быть довольно низкими – это зависит от требований к быстродействию и бюджету непосредственного заказчика данной системы.

2. РЕЗУЛЬТАТЫ РАСЧЕТОВ

Расчет производился на тестовых данных и модель еще будет адаптирована под задачу. Тестовыми данными выступали три набора данных: CLEVR CoGenT [27], Oscar train dataset [11] и Reciept-AVQA-2023. Результаты приведены в табл. 1. В качестве метрики используется точность ответов, т. к. в англоязычных источниках используют только ее, и реже – F-меру. Поэтому в первых экспериментах модели KVQA F-мера не была записана. Во всех последующих экспериментах на модели KVQA будет использована также и F-мера.

Oscar train dataset является небольшим улучшением набора VQA v2 [2, 3] и представлен в виде различных сцен в естественной среде.

CLEVR CoGenT является набором для проверки понимания моделью сложных взаимосвязей между объектами. Состоит из сцен с простыми фигурами. Вариант B отличается от A и обучающей выборки другим набором комбинаций цветов и форм фигур, из-за чего модель чаще генерирует ошибочные ответы.

Метод с тандемом для набора AVQA заметно улучшил точность ответов. Но задача набора AVQA требует только числовые ответы с точностью до тысячных, из-за чего метод многоклассовой классификации все еще выдает много неверных ответов. Для подобных задач необходимо поменять архитектуру кросс-модального модуля на интерполяцию, что не было сделано в рамках этой работы.

Перейдем к тестированию отдельной модели – классификации эмоций по аудио. Для ее тестирования было выбрано два набора данных – англоязычный и русскоязычный: CREMA-d [28] и RAMAS [29] соответственно.

Набор CREMA представляет собой 7.442 аудиофайла, на которых 91 актер с различными эмоциями произносят 12 фраз. Имеется так же и видео для кросс-модальных задач, что пригодится в дальнейших исследованиях в рамках этой работы.

В качестве признаков этих аудиофрагментов были извлечены их мел-кепстральные коэффициенты (MFCC). Наша сверточная модель показала следующие результаты: 82% точность на обучающей и 72% на тестовой выборках.

Набор RAMAS же не предоставляет возможность обучить на нем модель, ввиду малой размерности, не структурированности и отсутствия разметки. Поэтому, было принято решение использовать его как тестовый набор для уже обученной модели, а результаты проверять вручную.

Из-за ручной проверки результатов, точных цифр оценки нет, но было выявлено, что описанная выше модель, обученная на CREMA, плохо классифицирует отрывки набора RAMAS, ввиду отличий как в качестве записей, так и в отличии голосов актеров между этими двумя наборами. Поэтому, в данный момент идет изучение публичных моделей (как предобученных, так и обучаемых) для обработки голоса. Планируется применить более сложную модель, разработанную под широкий набор задач, и модифицировать ее для нашей классификации.

Все эксперименты над указанными моделями были проведены на мощностях ЦКП “Информатика” ФИЦ ИУ РАН. Были использованы следующие ресурсы Центра:

- 1x ЦП Intel Xeon Platinum 8160 (2.1 ГГц, 24 ядер, 48 потоков)
- ОЗУ 1536 ГБ
- 2x ГП Nvidia V100

Необходимо отметить, что подобные мощности требуются только на этапе подготовки набора данных и обучения моделей. Обученная модель может работать на намного более низких характеристиках хост-машины. Единственное ограничение – ОЗУ, ее нужно столько, чтобы в ней

можно было разместить несколько изображений (оперативные данные, загруженные в момент времени) и модель – от 2 ГБ.

3. ЗАКЛЮЧЕНИЕ

В работе представлен механизм мультимодального анализа текстовых и видеоданных, позволяющий распознать состояние эмоционального выгорания пользователя.

Изучен и подробно описан мультимодальный механизм VQA, а также особенности его реализации на примерах из других исследований.

Представлена собственная модель KVQA, протестированная на публичных наборах данных. Результаты для стандартной задачи VQA оправдали ожидания. Результаты на числовом множестве ответов неудовлетворительны, хоть и были улучшены. Требуется дальнейшее изучение проблемы. Механизм протестирован на мультиязычных наборах данных для классификации эмоций по аудио. Также протестирована модель распознавания отрицательного психоэмоционального состояния на основе аудиоданных.

Данные модели показывают хорошие результаты, и поэтому дальнейшая разработка системы распознавания психоэмоционального состояния человека будет основана на этих моделях.

Исследование выполнено в рамках научной программы Национального центра физики и математики, направление № 9 “Искусственный интеллект и большие данные в технических, промышленных, природных и социальных системах”.

Работа выполнялась с использованием инфраструктуры Центра коллективного пользования “Высокопроизводительные вычисления и большие данные” (ЦКП “Информатика”) ФИЦ ИУ РАН (г. Москва).

СПИСОК ЛИТЕРАТУРЫ

1. *Agrawal A., Lu J., Antol S., et al.* // Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015
2. *Zhang P., Goyal Y., Summers-Stay D., et al.* // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. P. 5014.
3. *Goyal Y., Khot T., Summers-Stay D., et al.* // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. P. 6904.
4. *Gordon D., Kembhavi A., Rastegari M., et al.* // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. P. 4089.
5. *Yan M., Li C.* https://github.com/alibaba/AliceMind
6. *Li L.H., Yatskar M., Yin D. et al.* https://arxiv.org/abs/1908.03557
7. *Das A., Datta S., Gkioxari G., et al.* // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. P. 1.
8. *Deng Y., Guo D., Guo X., et al.* http://arxiv.org/abs/2003.04641
9. *Batra D., Chang A.X., Chernova S., et al.* http://arxiv.org/abs/2011.01975
10. *Tan H., Bansal M.* // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. 2019.
11. *Li X., Yin X., Li C., et al.* http://arxiv.org/abs/2004.06165
12. *Zhang P., Li X., Hu X., et al.* // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. P. 5579.
13. *Yan M., Xu H., Li C., et al.* http://arxiv.org/abs/2111.08896
14. *Vaswani A., Shazeer N., Parmar N., et al.* http://arxiv.org/abs/1706.03762
15. *Devlin J., Chang M., Lee K., Toutanova K.* // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019. https://doi.org/10.18653/v1/N19-1423
16. *Wolf T., Debut L., Sanh V., et al.* // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. Stroudsburg, PA, USA: Association for Computational Linguistics, https://doi.org/10.2019. P. 38.
17. *Sennrich R., Haddow B., Birch A.* http://arxiv.org/abs/1508.07909
18. *Singh A., Goswami V., Natarajan V., et al.* https://github.com/facebookresearch/mmf
19. *Lin T., Maire M., Belongie S., et al.* http://arxiv.org/abs/1405.0312
20. *Kuznetsova A., Rom H., Alldrin N., et al.* http://arxiv.org/abs/1811.00982

21. *Shao S., Li Z., Zhang T., et al.* // In Proceedings of the IEEE international conference on computer vision. 2019. P. 8430.
22. *Krishna R., Zhu Y., Groth O., et al.* http://arxiv.org/abs/1602.07332
23. *Baek Y., Lee B., Han D., et al.* http://arxiv.org/abs/1904.01941
24. *Bautista D., Atienza R.* // ECCV 2022. Lecture Notes in Computer Science. Springer, Cham, 2022. P. 178. https://doi.org/10.1007/978-3-031-19815-1_11
25. *Chernykh V., Prikhodko P.* Emotion recognition from speech with recurrent neural networks // https://arxiv.org/abs/1701.08071
26. *Lin W.C., Emmanouilidou D.* Toxic Speech and Speech Emotions: Investigations of Audio-based Modeling and Intercorrelations //2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022. P. 115.
27. *Johnson J., Hariharan B., Maaten L., et al.* CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning // CVPR 2016
28. *Cao H., Cooper D.G., Keutmann, M.K., et al.* // IEEE transactions on affective computing, 2014. P. 377. https://doi.org/10.1109/TAFFC.2014.2336244
29. *Perepelkina O., Kazimirova E., Konstantinova M.* // ResearchGate preprint 03.2018 https://doi.org/10.7287/peerj.preprints.26688v1

PROTOTYPE SYSTEM FOR RECOGNIZING HUMAN FATIGUE STATES USING VIDEO, AUDIO, AND TEXT DATA

D. A. Weizenfeld[1,2], G. A. Kiselev[1,#], Y. S. Korovin[3], and S. V. Makov[4]

[1]Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russia

[2]Patrice Lumumba Peoples' Friendship University of Russia, Moscow, Russia

[3]Research Institute of Multiprocessor Computation Systems, Taganrog, Russia

[4]Institute of Services Sector and Entrepreneurship, Shakhty branch, Don State Technical University, Shakhty, Russia

[#]e-mail: kiselev@isa.ru

A prototype system utilizing video, audio, and text data for recognizing states of fatigue and reduced human performance is described. For this purpose, the task of Visual Question Answering (VQA) has also been studied and elaborately outlined, along with features of its implementation based on examples from another research. Experiments have been conducted on datasets with a wide range of tasks: the standard VQA task on the VQA v2 dataset, complex scenarios on CLEVR CoGenT, and analysis of cash receipts on Receipt-AVQA-2023.