УДК 004.891

ОБУЧЕНИЕ ВОПРОСНО-ОТВЕТНОЙ НЕЙРОСЕТЕВОЙ МОДЕЛИ НА БАЗЕ АРХИТЕКТУРЫ МОДЕЛИ LLAVA 1.5 С ЭНКОДЕРОМ SAIGA MISTRAL 7B И АЛГОРИТМА НИЗКОРАНГОВОЙ АДАПТАЦИИ LORA

Л.А. Демидова
д.т.н., профессор кафедры корпоративных информационных систем Института информационных технологий МИРЭА – Российского технологического университета, Москва, Россия; orcid.org/0000-0003-4516-3746, e-mail: demidova.liliya@gmail.com

Н.А. Морошкин
аспирант кафедры корпоративных информационных систем Института информационных технологий МИРЭА – Российского технологического университета, Москва, Россия; orcid.org/0009-0002-8787-2452, e-mail: moroshkin@mirea.ru

Л.А. Демидова

Окончила Московский государственный университет им. М.В Ломоносова (МГУ). Область научных интересов связана с разработкой методов и алгоритмов интеллектуального анализа данных.

Н.А. Морошкин

Окончил МИРЭА – Российский технологический университет. Область научных интересов связана с разработкой вопросно-ответных нейросетевых моделей компьютерного зрения и методов оптимизации, конвертации нейросетевых моделей.

Аннотация. В данной работе реализован алгоритм обучения визуальной вопросно-ответной нейросетевой модели на базе архитектуры нейросетевой модели *LLaVA* 1.5 с использованием текстового энкодера *Mistral* 7*b*, позволяющего улучшить результаты работы модели в задаче визуального вопросно-ответного моделирования и алгоритма низкоранговой адаптации *LoRA,* позволяющего ускорить процесс обучения модели.

Показано, что задача вопросно-ответного моделирования может быть решена нейросетевыми моделями с использованием больших языковых моделей, описана методика ускорения обучения таких моделей. Проведена оценка эффективности разработанной модели и показаны общие аспекты обучения вопросно-ответных моделей на наиболее популярных вопросно-ответных наборах данных.

Ключевые слова: Визуальное вопросно-ответное моделирование, большие языковые модели, алгоритм низкоранговой адаптации *LoRA*, нейросетевая модель *LLaVA* 1.5, *Saiga*/*Mistral* 7*b*.

Введение. Вопросно-ответное моделирование (*BOM*, *Question Answering*, *QA*) давно и широко применяется при решении различных прикладных задач [1]. Однако, начиная с 2017 года [2], специалисты-разработчики начали проявлять интерес к синтезу вопросно-ответных систем (ВОС) с использованием технологий искусственного интеллекта, а именно – с применением нейросетевого моделирования. При этом особое внимание стало уделяться визуальному вопросно-ответному моделированию (*ВВОМ*, *Visual Question*

Answering, *VQA*), которое обычно предполагает решение мультимодальной задачи обработки больших данных.

Преимущество такого подхода заключается в возможности адаптации ВОС к решению множества задач визуального анализа данных, например, к детектированию объектов на визуальных данных, классификации визуальных объектов. Такие подходы реализуются с помощью создания текстовых запросов к нейросетевым моделям [1].

В последних работах [3,4], посвященных обработке текстовых данных, наиболее популярное решением является использование больших языковых моделей. Подобные модели позволяют решать множество задач обработки естественного языка [4]. Задача визуального вопросно-ответного моделирования – это мультимодальная залача, так как она содержит в себе объединение подходов к обработке визуальных и текстовых данных. Вопрос применения больших языковых моделей в качестве текстовых энкодеров в решениях визуального вопросно-ответного моделирования исследован в меньшей степени, однако исследователи нейросетевых технологий начинают использовать инновации больших языковых моделей для построения больших мультимодальных моделей [5].

Визуальная вопросно-ответная нейросетевая модель *LLaVA* 1.5. Нейросетевая модель *LLaVA* 1.5 – это большая мультимодальная модель, решающая задачу вопросно-ответного моделирования с использованием большой языковой модели в качестве энокера текстовых данных [5]. Преимущества данной архитектуры заключаются в открытом исходном коде, в возможности использования множества известных больших языковых моделей, а также – в возможности решения задачи распознавания символов на изображениях.

Архитектура модели представлена на рисунке 1. Ядро модели составляют большая языковая модель *Vicuna-7b* (с 7 миллиардами параметров) и визуальный энкодер *CLIP ViT-L*/14 с 428 миллионами параметров.

Ядро алгоритма работы модели *LLaVA* состоит из соединения двух энкодеров – визуального и текстового. Для входной многомерной матрица изображения X_v используется визуальный энкодер (в оригинальной архитектуре версии 1.5 нейросетевая модель *CLIP ViT-L*/14), который обрабатывает изображение, получая векторное представление Z_v . Затем векторное представление Z_v передается в многослойный перцептрон, размерность выходного слоя которого совпадает с размерностью входного слоя большой языковой модели (в оригинальной архитектуре версии 1.5 используется большая языковая модель *Vicuna*), получая векторное представление изображения H_v . Текстовый запрос к модели X_q (вопрос к вопросно-ответной системе) токенизируется с помощью алгоритма токенизации. Под токенизацией в контексте этой работы имеется ввиду алгоритм разделения текста на элементарные составляющие (символы или слова). В результате формируется токенизированное представление H_q, размерность которого равна размерности входного слоя большой языковой модели. Наконец, векторные представления изображения H_v и текстового запроса H_q передаются в языковую модель. В результате формируется текстовый ответ X_a.

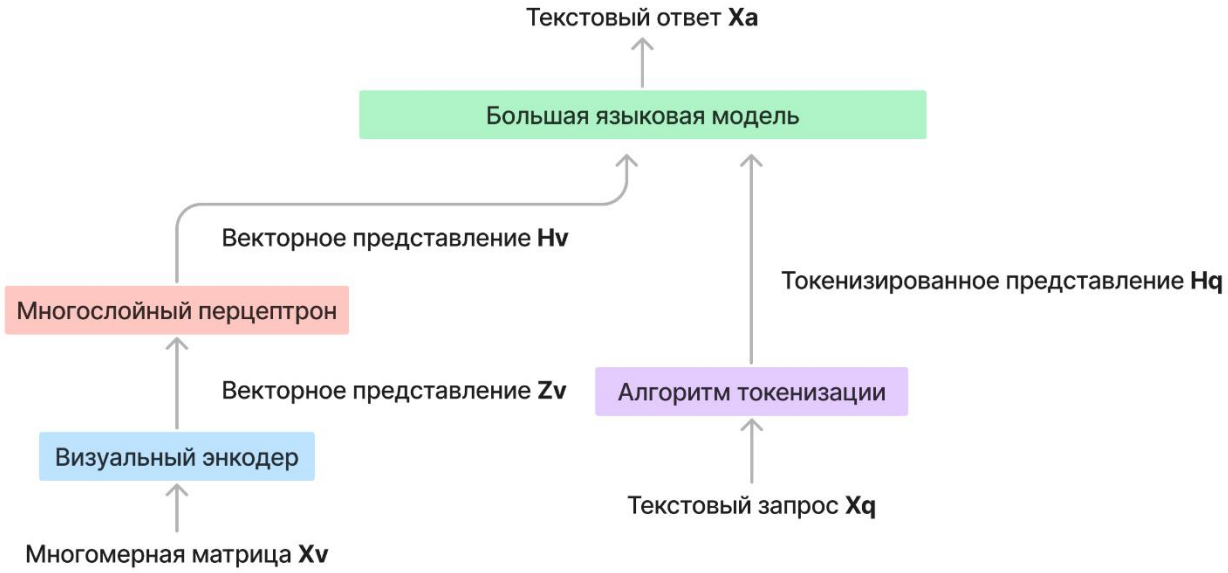

Рисунок 1. Архитектура вопросно-ответной нейросетевой модели *LLaVA* 1.5

Особенность модели *LLaVA* заключается в возможности обучения модели без использования больших наборов данных по сравнению с другими архитектурами [5].

Изменив стандартный подход к формированию запроса [5], модель обладает лучше обобщающей способностью и способна отвечать развернуто на запросы. Стандартный подход состоял в том, что модели задавался один или несколько вопросов к одному контексту (в задаче *BBOM* контекстом является входное изображение). В модели *LLaVA* было предложено использовать множественные вопросы к множественным контекстам. Тем самым, авторы улучшали способность к обобщению у модели.

Кроме того, изменен подход к обучению вопросно-ответной модели: в вопрос к ВОС добавлены инструкции, которые указывают большой мультимодальной модели на необходимость распознавания символов на изображениях для более точного ответа. Оценки точности модели *LLaVA* и некоторых других *SOTA* моделей представлены в таблице 1. В таблице представлены наиболее распространенные наборы данных для обучения и тестирования моделей *BBOM*.

Таблица 1. Оценки точности *SOTA* моделей при решении задачи *VQA* на разных наборах данных [6]

Имя модели	Набор данных *VQAv2*, метрика *accuracy* [5]	Набор данных *GQA*, метрика *accuracy* [5]	Набор данных *POPE*, метрика *accuracy* [5]
LLaVA 1.5	0,8	0,63	0,85
InstructBLIP	0,48	0,49	0,79
BLIP-2	0,15	0,02	0,86

Авторы модели *LLaVA* уделили большое внимание возможности масштабирования входного изображения, улучшив обобщающую способность визуального энкодера. В следующей версии 1.6 идея масштабирования входного изображения была расширена, что позволило улучшить метрические показатели модели (*f*-мера [6], точность).

В качестве наборов обучающих данных для модели *LLaVA* 1.5 можно использовать стандартные наборы данных для *BBOM*, наборы данных визуального регионального представления (*REVIVE*, *Revive Dataset*) и наборы данных для решения задачи распознавания символов на изображениях (*OCR*, *Optical Character Recognition*).

Модель *LLaVA* характеризуется также способностью работать с несколькими языками. Несмотря на то, что модель изначально не была адаптирована под многоязыковую задачу, она способна обрабатывать запросы на не родных языках. Однако качество таких ответов хуже, чем у аналогичных моделей. Главные особенности модели, такие как региональное представление и возможность распознавать символы на изображениях, теряются из-за невозможности большой языковой модели обработать векторные представление неродных языков.

Текстовый энкодер *Saiga*/*Mistral* 7*b*. В данной работе предлагается обучить модель *LLaVA* на оригинальном наборе данных, заменив текстовый энкодер на модель *Saiga*/*Mistral* 7*b*, обученную на наборе данных на русском языке. Модель *Saiga*/*Mistral* 7*b* использует внимание к групповым запросам для ускорения работы и алгоритм скользящего окна внимания для обработки последовательности произвольной длины [7]. Учитывая, что оригинальная модель *Mistral* была обучена с применением групповых запросов и то, что модель *LLaVA* использует аналогичные данные для валидации и аннотации, можно сделать предположение о том, что модель *Saiga*/*Mistral* 7*b* способна заменить модель *Vicuna* в роли текстового энкодера в архитектуре *LLaVA*. Оценки точности модели *Mistral* 7*b* и её некоторых аналогов представлены в таблице 2. Учитывая специфику вопросно-ответного моделирования, сравнение больших текстовых моделей представлено в виде MT-Bench оценки, основанной на рейтинговой системе. Эта оценка показывает, насколько модель способна вести многоходовый диалог, начиная с одного из 80 подготовленных вопросов [7].

Таблица 2. Оценки точности больших языковых моделей для вопросно-ответного моделирования [7]

Имя модели	Набор данных *MT-Bench* [7]
Mistral 7*b Instruct*	6,84
Llama 2 7*b*	6,65
Vicuna 13*b*	6,17

Набор данных *Saiga*, использующийся для обучения модели *Saiga*/*Mistral* 7*b,* основан на принципе работы большой языковой модели *Baize* [8] как конвейера для генерации многоходового корпуса текстовых данных. Данный подход адаптирован под решение задачи вопросно-ответного моделирования [9] и демонстрирует хорошую обобщающую способность для ведения многоходового диалога.

Обучение вопросно-ответной модели с использованием алгоритма низкоранговой адаптации *LoRA*. При обучении вышеописанной визуальной вопросно-ответной модели предлагается использовать алгоритм низкоранговой адаптации *LoRA*, который в свою очередь применяется в процессе обучения оригинальной модели *LLaVA* 1.5, а также в некоторых других моделях для улучшения производительности обработки разных доменов корпуса текстов [6].

В основу алгоритма *LoRA* [5] закладывается разделение матрицы весов модели на две матрицы меньшего размера. Эти матрицы определяют новые весовые коэфициенты нейросетевой модели. Затем модель с новыми весовыми коэфициентами обучается стандартным методом [5]. Такой подход позволяет использовать меньше вычислительных ресурсов, так как две меньшие матрицы требуют ресурсов на хранение и обработку меньше, чем одна большая, составленная из двух малых. Алгоритм *LoRA* получил широкое распространение в исследованиях больших языковых моделей [7].

При проектировании архитектуры решения было необходимо внести изменения в оригинальную архитектуру, поскольку размер векторного представления оригинальной модели *Vicuna* 13*b* и модели *Saiga*/*Mistral* 7*b* отличаются: они равны 5120 выходных

нейронов и 4096 выходных нейронов соответственно. Для дальнейшего обучения необходимо модифицировать архитектуру многослойного персептрона, изменив количество выходных нейронов с 5120 на 4096.

При обучении наборы данных не изменялись, хотя в них не содержалась ответов на русском языке. Однако исследователи обнаружили [8] способность моноязыковых больших моделей обрабатывать многоязыковые задачи. Для сравнения двух моделей – *LLaVA* с энкодером *Saiga*/*Mistral* 7b (*LLaVA-Saiga*/*Mistral* 7b) и оригинальной *LLaVA* был применен следующий подход.

1 Перевести на русский язык три набора данных, представленных в таблице 1 (*VQAv*2, *GQA*, *POPE*).

2 Получить ответы модели *LLaVA-Saiga*/*Mistral* 7*b* на вопросы из входных наборов данных.

3 Перевести ответы модели на английский язык и сравнить с ответами оригинальной модели *LLaVA* 1.5.

Такой подход обусловлен отсутствием визуального вопросно-ответного набора данных для обучения на русском языке. В эксперименте было решено не использовать синтетические наборы данных во избежание потери способности обобщения, так как модель *Saiga*/*Mistral* 7*b* обучалась на синтетических данных. Оценки точности двух моделей представлены в таблице 3. Жирным шрифтом выделены оценки точности для модели *LLaVA* с текстовым энкодером *Saiga*/*Mistral* 7*b.*

Обучение длилось одну эпоху, как и предлагают авторы модели *LLaVA* [5], при этом был использован алгоритм низкоранговой адаптации *LoRA*. Шаг обучения, функции потерь и основные параметры алгоритмы *LoRA* при обучении не изменялись по сравнению с предложенными значениями авторами в оригинальной статье [5]. Во время проведения эксперимента, было отмечено, что весовые коэффициенты полученные с помощью алгоритмы *LoRA*, требуют около 80 Гб оперативной памяти графического процессора. Обучение проходило на одном графическом ускорителе с тензорными ядрами *NVIDIA A*100. Одна эпоха продлилась приблизительно одну неделю. Во время подготовки к процессу обучения была также замечена особенность модели *LLaVA*: обучить оригинальную модель удалось только на графическом ускорителе архитектуры *Ampere*, на других архитектурах графических ускорителей обучение оказалось невозможным из-за несовместимости рассматриваемой модели с архитектурой *CUDA*.

Таблица 3. Оценки точности оригинальной модели *LLaVA* 1.5 и модели *LLaVA* 1.5 с энкодером *Saiga*/*Mistral* 7*b*

Имя модели	Набор данных *VQAv*2, метрика *accuracy* [5]	Набор данных *GQA*, метрика *accuracy* [5]	Набор данных *POPE*, метрика *accuracy* [5]
LLaVA 1.5	0,8	0,63	0,85
LLaVA-Saiga/Mistral	**0,72**	**0,53**	**0,79**

Результаты, представленные в таблице 1, показывают, что изменение текстового энкодера в архитектуре на другом языке ухудшают точность модели. Однако, учитывая, что модель *LLaVA-Saiga*/*Mistral* 7*b* обучалась на наборе данных на другом языке, возможно, точность удастся улучшить на наборе реальных данных на русском языке.

Однако основные свойства и особенности модели сохранились, как было отмечено во время проведения ручного тестирования модели, вопросно-ответная система сохранила свою возможность вести многоходовые диалоги, отвечать на несколько вопросов заданных по одному изображению. Сохранилась возможность обобщения визуальных данных. Модель стала лучше различать кириллические рукописные символы на

изображениях, что позволяет более точно решать задачу распознавания символов на изображениях. Исходя из этих наблюдений, можно сделать вывод, что для улучшения точности модели и получения лучших значений показателя точности модели, чем у оригинальной модели *LLaVA*, требуется провести больше одной эпохи и подготовить набор данных для *BBOM* на русском языке. Возможно, его удастся синтезировать, аналогично набору данных *Saiga*, который использовался для обучения модели *Saiga*/*Mistral* 7*b*.

Заключение. Проведенный эксперимент показывает возможность обучения вопросно-ответной модели на основе архитектуры *LLaVA* методом замены энкодера. Учитывая результаты сравнения, можно сделать вывод, что современные большие языковые модели, способные решать задачи на языках, отсутствующих в обучаемой выборке, способны решать задачу векторного представления в вопросно-ответных моделях.

Список литературы

[1] Демидова Л.А., Морошкин Н.А. Аспекты разработки архитектуры вопросно-ответной системы для обработки больших данных на основе нейросетевого моделирования // Вестник Рязанского государственного радиотехнического университета. 2023. No 86. С. 55–69.

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh VQA: Visual Question Answering // International Conference on Computer Vision (ICCV) 2015, 2015. pp. 1-13.

[3] Husein Zolkepli, Aisyah Razak, Kamarul Adha, Ariff Nazhan Multi-Lingual Malaysian Embedding: Leveraging Large Language Models for Semantic Representations // arXiv.org, 2024, DOI:10.48550/arXiv.2402.03053

[4] Sophie Xhonneux, David Dobre, Jian Tang, Gauthier Gidel, Dhanya Sridhar In-Context Learning Can Re-learn Forbidden Tasks // arXiv.org, 2024, DOI:10.48550/arXiv.2402.05723

[5] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee Visual Instruction Tuning // arXiv.org, 2023, DOI:10.48550/arXiv.2304.08485

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee Improved Baselines with Visual Instruction Tuning // arXiv.org, 2023, DOI:10.48550/arXiv.2310.03744

[7] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed Mistral 7B // arXiv.org, 2023, DOI:10.48550/arXiv.2310.06825

[8] Mikhail Tikhomirov, Daniil Chernyshev Impact of Tokenization on LLaMa Russian Adaptation // arXiv.org, 2023, DOI:10.48550/arXiv.2312.02598

Авторский вклад

Демидова Лилия Анатольевна – руководство исследованием по оценке точности модели, постановка задачи исследования, описание общей постановки задачи вопросно-ответного моделирования

Морошкин Никита Андреевич – Реализация процесса обучения модели LLaVA с текстовым энкодером Saiga/Mistral 7b, сравнение рассматриваемой модели с оригинальной моделью LLava, анализ полученных результатов

TRAINING A QUESTION AND ANSWERING NEURAL NETWORK MODEL BASED ON LLAVA 1.5 MODEL ARCHITECTURE WITH SAIGA MISTRAL 7B ENCODER AND LOW-RANK ADAPTATION ALGORITHM LORA

L.A. Demidova
Doctor of Technical Sciences, Professor of the Department of Corporate Information Systems, Institute of Information Technologies MIREA - Russian Technological University, Moscow, Russia;

N.A. Moroshkin
postgraduate student of the Department of Corporate Information Systems, Institute of Information Technologies MIREA - Russian Technological University, Moscow, Russia; orcid.org/0009-0002-8787-2452,

Annotation. In this work, a training algorithm for a visual question-answer neural network model is implemented based on the structure of the LLaVA 1.5 neural network model using the Mistral 7b text encoder, which allows improving the model's performance in the task of visual question-answer modeling and the LoRA low-rank adaptation algorithm, which speeds up the model learning process .

It is shown that the problem of question-answer modeling can be solved by neural network models using large language models, and a technique for accelerating the training of such models is described. The effectiveness of the algorithmic model is assessed and general aspects of training question-answer models on the most popular sets of question-answer data are shown.

Keywords: Visual question-answer modeling, large language models, low-rank adaptation LoRA, neural network model LLaVA 1.5, Saiga/Mistral 7b.