Сравнение способов до-обучения моделей ИИ

Скуредин Сергей

Июнь 2025

Резюме

Данная работа направлена на рассмотрение и оценку различных средств до-обучения нейросетевых моделей в области NLP при решении задач логического вывода. В работе использовалась современная пред-обученная модель Qwen3 с набором данных TERRa, входящим в бенчмарк «Russian SuperGLUE». Результаты эксперимента показывают, что затрачиваемое время на обучение моделей может быть значительно менее важно, чем выбор подходящих способов обучения. Репозиторий с файлами эксперимента расположены по адресу https://github.com/CACTEP/NLP.fpmidpo.mipt

1 Введение

Задача классификация пары предложений (Recognizing Textual Entailment) интересна сама собой в силу её основополагающего свойства для реализации одной из ключевых миссий дисциплины «обработка естественного языка» (natural language processing). Эта миссия заключена в создании возможности свободного диалога между человеком и машиной. В свою очередь, логические связи, естественно понимаемые человеком, достаточно сложно воспринимаются машиной. С другой стороны, стремительный прогресс в технологиях создания и обучения нейросетевых моделей в области NLP создает непреодолимые сложности для индивидуальных исследователей, связанные с гонкой в обеспечении вычислительных возможностей, а также величиной количества параметров, используемых для поиска и тестирование гипотез.

Данная работа является попыткой ответить на определенные выше два фактора и определить возможности и области дальнейших исследований в области NLP.

1.1 Команда

Над задачей трудился обучающийся по программе курса «Natural Language Processing» в МФТИ ФПМИ Скуредин Сергей Андреевич, включая подготовку техническую часть и подготовку отчета.

2 Работы других авторов

В настоящей работе использованы следующие труды других исследователей:

Статья «RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark» от 2020 г., в которой авторы представили, разработанный ими, бенчмарк оценки общего понимания русского языка, построенный по аналогичным принципам бенчмарка оценки общего понимания английского языка – SuperGLUE (Shavrina et al., 2020). Авторы проделали огромную работы по выбору, проверки и сведению наборов данных, которые могут быть использованы для обучения и тестирования способности моделей в решении задач типа логического вывода, здравого смысла и др.

Статья «Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models» от 2025 г, в которой авторы представили серию моделей Qwen3 Embedding, продолжающую линейку мультиязычных моделей Qwen (Zhang et al., 2025). Примечательным является то, что не смотря на относительно небольшие по современным меркам размеры моделей, авторам удалось достигнуть хороших показателей качества, даже на уровне других моделей с большим числом параметров.

3 Описание модели

Модель, используемая в настоящей работе, описывается следующими компонентами:

1) Базовая модель – существующая пред-обученная модель.

2) Доработанная модель – изменения поверх базовой модели, которые демонстрируют наилучший результат.

В качестве базовой модели использовалась современная пред-обученная Qwen3 Embedding-0.6B. Базовая модель была построена поверх фундаментальной модели Qwen3-0.6B-Base и обучена в два этапа: первый – обучение без учителя на синтезированных данных; второй – обучение с учителем на высококачественных данных (Zhang et al., 2025). Архитектура базовой модели представлена на рис.1.

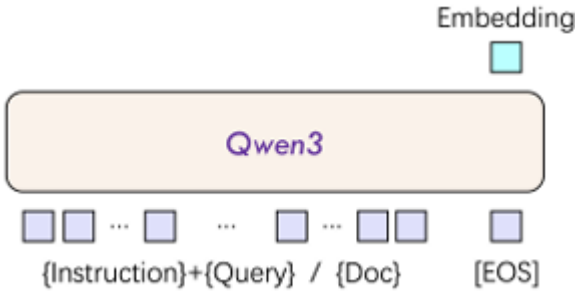

Рисунок 1: Архитектура базовой модели Qwen3 Embedding-0.6B.

Характеристики базовой модели представлены в табл. 1

Кол-во параметров	Кол-во слоев	Максимальная длина последовательности	Размер выходного векторного представления	Размер файла с моделью
0.6 млрд.	28	32k токенов	1024	1.19 Гб

Таблица 1: Характеристики базовой модели Qwen3 Embedding-0.6B.

Доработанная модель включает базовую пред-обученную модель, добавленный выходной слой для бинарной классификации, а также добавленные параметры к выборочным слоям – адаптеры LoRA.

Характеристики доработанной модели представлены в табл. 2

Общее кол-во параметров	Кол-во добавленных параметров	Слои, к которым добавлены параметры	Размер выходного векторного представления	Настройки адаптера LoRA
596,928,512	1,148,928	«q_proj», «v_proj»	2	rank – 8, alpha – 32, dropout – 0.1, task type – SEQ_CLS

Таблица 2: Характеристики базовой модели Qwen3 Embedding-0.6B.

4 Описание набора данных

Для настоящей применена часть популярного в России набора данных, используемого для тестирования продвинутых интеллектуальных способностей: логика, формирование выводов, рассуждения на основе здравого смысла (Shavrina et al., 2020). В частности, для эксперимента использован набор данных «TERRa», направленный на поиск решения задачи типа классификации пары предложений (Recognizing Textual Entailment) – распознавание корректности вывода (логического следствия, связи) по отношении к заданному предложению (первичное условия) (Natural Language Inference).

Характеристики набора данных приведены в таб. 3.

Кол-во возможных вариантов правильного ответа	Соотношение положительных и отрицательных ответов (обучение / валидация)	Кол-во признаков в одной единице данных	Количество единиц данных (обучение / валидация)	Размер файлов с данными
2	0.52 к 0.48 / 0.50 к 0.50	4	2616 / 307	3.44 Мб

Таблица 3: Характеристики набора данных «TERRa».

Пример используемого наборы данных имеет следующий вид:

{

"premise": "Музей, расположенный в Королевских воротах, меняет экспозицию. На смену выставке, рассказывающей об истории ворот и их реставрации, придет «Аптека трех королей». Как рассказали в музее, посетители попадут в традиционный интерьер аптеки.",

"hypothesis": "Музей закроется навсегда.",

"idx": 10,

"label": 1

}

Набор данных собирался путем фильтрации крупного корпуса Taiga (Шаврина и Шаповалова, 2017) по правилу – в первом предложении должен быть осмысляемые глаголы, а второе предложение должно иметь связь с первым предложением наличием осмысленного глагола, которое кратко излагает суть действий первого предложения. При этом подбирались предложения с часто используемыми словами – количество на миллион.

5 Описание эксперимента

5.1 Метрики качества задачи

Для оценки результатов эксперимента применялись метрики, направленные на измерение качества модели в решении задачи, а также связанные с производительность обучения модели. Качество модели оценивается точностью (Accuracy), а также уровнем ошибок (Loss). Производительность обучения модели оценивается затраченным временем и количеством обучаемых параметров.

5.2 Параметры и дизайн эксперимента

В качестве технического обеспечения эксперимента использовалась рядовой графический ускоритель Nvidia GeForce RTX 4070 Ti Super, имеющий следующие важные характеристики:

Объем памяти – 16 376 Мб;
Пропускная способность памяти – 672.3 Гб/с;
Количество вычислительных ядер CUDA – 8448;
Теоретическая производительность – 44.61 TFLOPS (FP 16/32).

Эксперимент имеет следующие ключевые параметры:
Количество эпох обучений и итераций в каждой эпохе – 4 и 109, значения которые подобраны эмпирически, чтобы одновременно установить равные условия циклов обучения и провести эксперимент в разумные сроки;
Размер батча – 24, значение которое подобрано эмпирически, с учетом предположения о возможной большой разнице в длинах между единицами набора данных и влиянии этого на длительность расчета одной итерации в эпохе обучения.

Следует отметить, что память графического ускорителя, с учетом помех, вызванных работой операционной системы, при каждом запуске обучения занимала практически всю память (16 250 Мб – 16 300 Мб). Однако, при использовании LoRA задействованная память заметно возрастала на несколько десятков Мб (16 300 Мб – 16 350 Мб).

Эксперимент описывается следующей последовательностью действий:
Обработка обучающих данных для возможности их использования в непосредственном обучении.
Применение техники 1 – до-настройка модели (Fine-tuning): полное обучение пред-обученной базовой модели, включая добавленный слой для решения задачи бинарной классификации. Данная техника в силу своей очевидности является целевым минимальным уровнем (baseline) для проведения сравнений и оценки результатов.
Применение техники 2 – обучение только добавленного слоя (Linear Probing): обучения только для добавленного слоя, а параметры пред-обученной базовой модели были установлены в неизменяемый режим.
Применение техники 3 – обучение добавленных параметров к выборочным слоям (LoRA): обучение только добавленных LoRA-параметров к следующим слоям пред-обученной базовой модели:
выбранные слои: "q_proj", "v_proj";
другие параметры: rank – 8, alpha – 32, dropout – 0.1, task type – SEQ_CLS.
Применение техники 4 – использование пред-обученного добавленного слоя бинарной классификации (результат техники обучения только добавленного слоя) и полной до-настройки модели (Improving fine-tuning).
Применение техники 5 – использованию пред-обученного добавленного слоя (результат техники обучения только добавленного слоя) и обучения добавленных параметров к выборочным слоям с аналогичными параметрами:
выбранные слои: "q_proj", "v_proj";
другие параметры: rank – 8, alpha – 32, dropout – 0.1, task type – SEQ_CLS.

6 Результаты

По результатам эксперимента получен ряд значений оцениваемых показателей (таб. 4).

Оцениваемая техника до-обучения, рисунок с графиками обучения	Точность (выше лучше)	Уровень ошибки (ниже лучше)	Количество обучаемых параметров	Затраченное время
Техника 1, рис. 2 (baseline)	0.557	0.765	595,779,584	6h 9min 31s
Техника 2, рис. 3	0.537	0.861	2,048	4h 38min 55s
Техника 3, рис. 4	0.629	0.639	1,148,928	3h 23min 8s
Техника 4, рис. 5	0.606	0.676	595,779,584	6h 13min 50s
Техника 5, рис. 6	0.593	0.665	1,148,928	2h 32min 5s

Таблица 4: Результаты эксперимента.

Подробные графики метрик точности и уровня ошибки приведены в приложении.

7 Обсуждение

Данные результатов эксперимента свидетельствуют о том, что большее время, затрачиваемое на обучение модели, не всегда однозначно положительно сказываются на качестве решении поставленной задачи (допущение: возможно, автор работы выбрал не лучшим образом подходящую пред-обученную модель для решения выбранной задачи). Выбор техники, применяемой для до-обучения модели представляется намного более существенным.

Так, например, применение техники 3 представляется более перспективным если использовать больше временных ресурсов: линии тренда изменения показателей уровня ошибки и точности показывают более желательные направления (рис. 4). При этом, чуть более чем за три часа двадцать минут эта техника 3 показала наилучшее качество модели.

Напротив, ожидалось, что использование техники 5 будет сильнее техники 3 в связи с тем, что используется пред-обученный добавленный слой бинарной классификации, полученный в результате техники 2, который сможет значительно повысить качество за схожее с техникой 3 время. Но, результаты показывают противоположную ситуацию: время обучения значительно сократилось до двух с половиной часов (без учета временных затрат на предварительное обучение добавленного слоя бинарной классификации – времени техники 2), и качество модели оказалось ниже – почти на четыре сотых по точности и почти три сотых по уровню ошибки.

В свою очередь техники 1, 2 и 4 представляются мало перспективными с точки зрения временных затрат. Но у техники 4 графики обучения (рис. 5) внушают надежду, что с наличием больших вычислительных ресурсов, способных нивелировать временные затраты, обучение может привести к хорошему качеству.

В общем складывается устойчивое предположение, что автор ещё не имеет достаточного понимания или ресурсов (временных, технических), чтобы получить такие результаты, которые свидетельствовали бы о значительном прогрессе человечества в вопросе реализации миссии по обеспечению свободного диалога между человеком и машиной. Для последующих исследований рекомендуется как минимум тестирование других базовых пред-обученных моделей (включая разные вариации одной модели), а также, что, представляется намного сильнее, проводить эксперименты с разными способами (техниками) формирования и обучения моделей для решения задач обработки естественного языка, например, использовать технику обучения добавленных параметров на всех слоях, или искать принципиально другие алгоритмы.

Ссылки

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, Jingren Zhou.
Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, 2025.
URL https://arxiv.org/html/2506.05176v3

Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, Andrey Evlampiev.
RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark, 2020.
URL https://russiansuperglue.com/ru/download/main_article

Приложение

Графики метрик точности и уровня ошибки вариаций эксперимента

Техника 1 (baseline)

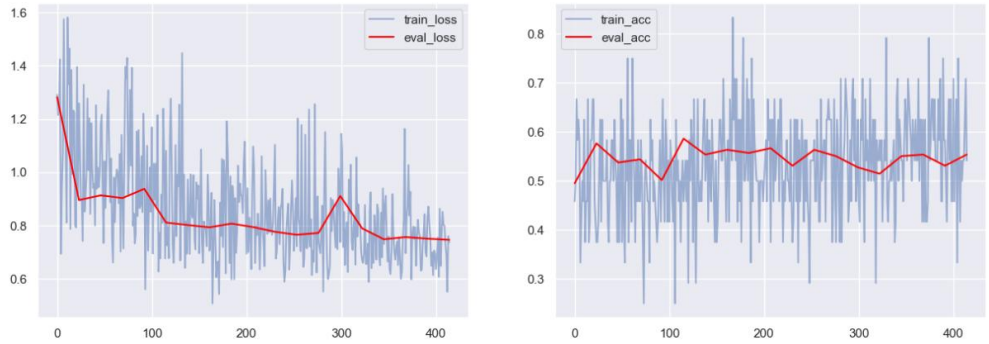

Рисунок 2: Графики изменения уровня ошибки и точности при обучении базовой пред-обученной модели и добавленного слоя бинарной классификации.

Техника 2

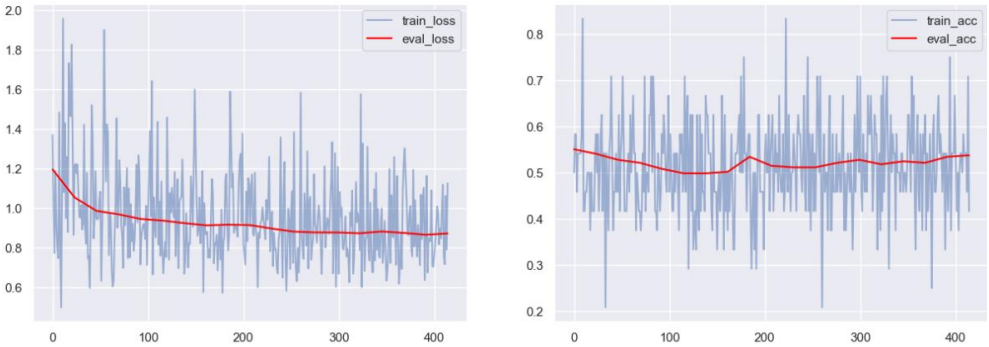

Рисунок 3: Графики изменения уровня ошибки и точности при обучении только добавленного слоя бинарной классификации; параметры базовой пред-обученной модели заморожены.

Техника 3

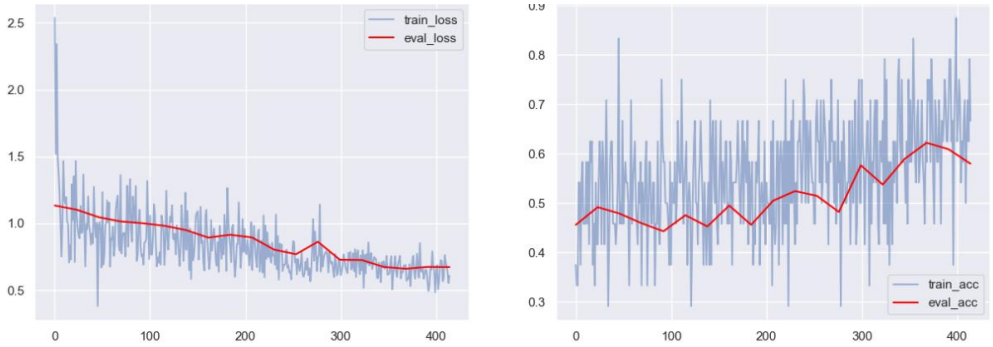

Рисунок 4: Графики изменения уровня ошибки и точности при обучении добавленных параметров к выборочным слоям базовой пред-обученной модели; без обучения дополнительного слоя бинарной классификации.

Техника 4

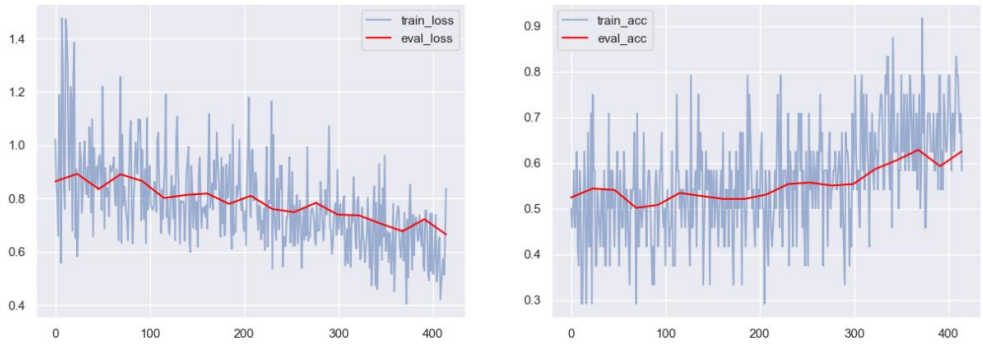

Рисунок 5: Графики изменения уровня ошибки и точности при обучении базовой пред-обученной модели и пред-обученного добавленного слоя бинарной классификации.

Техника 5

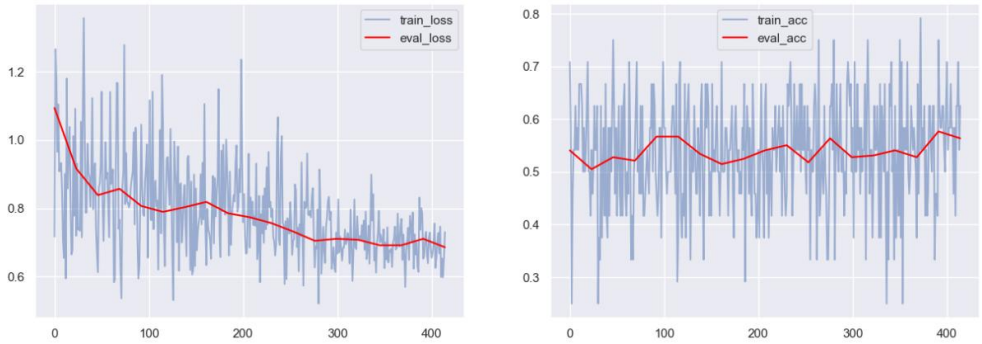

Рисунок 6: Графики изменения уровня ошибки и точности при обучении добавленных параметров к выборочным слоям базовой пред-обученной модели, у которой пред-обучен добавленный слой бинарной классификации.