УДК 004.4

ЯЗЫКОВАЯ МОДЕЛЬ ГЕНЕРАТИВНОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА FALCON LANGUAGE MODEL

Пронин А.А., студент гр. ИСт-222, II курс
Лакман А.А., студент гр. ИСт-222, II курс
Осипов Н.А., студент гр. ИАб-221, II курс
Бояринцев А.А., студент гр. ИАб-221, II курс
Герасимов В.М., студент гр. ИАб-221, II курс
Научный руководитель: Семенова О.С., к.т.н., доцент
ФГБОУ ВО «Кузбасский государственный технический университет им. Т.Ф. Горбачева»
г. Кемерово

Когда говорят о генеративных моделях искусственного интеллекта, подразумевают новое поколение моделей глубокого обучения, называемых базовыми моделями. Базовые модели (рис. 1) – это предварительно обученные модели искусственного интеллекта, которые можно настроить для конкретных задач.

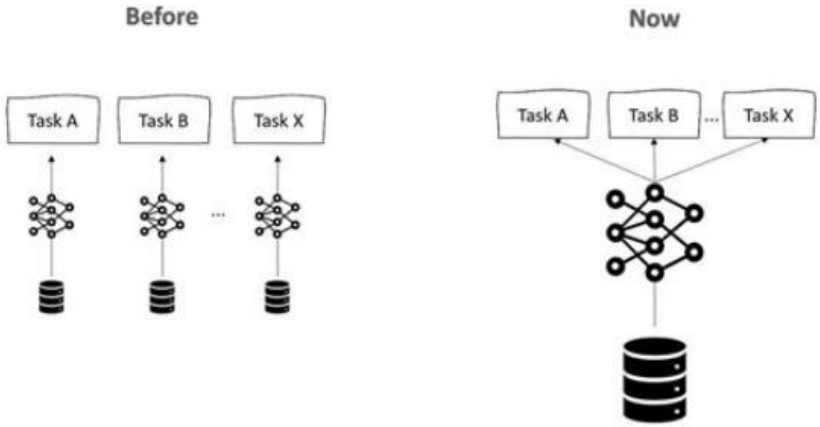

Рисунок 1 – Базовые модели

В случае ChatGPT и подобных моделей говорят о больших языковых моделях (LLM), подмножестве базовых моделей, специально разработанных для задач обработки естественного языка. Такие модели, как GPT-4, являются примерами LLM, которые могут генерировать текст, подобный человеческому, отвечать на вопросы, переводить языки и многое другое.

LLM характеризуются огромными наборами обучающих выборок и параметров. Для примера, GPT-3 был обучен почти на 500 миллиардах токенов и имеет 175 миллиардов параметров. Однако модели с таким большим количеством параметров являются «тяжелыми» как на этапе обучения, так и на этапе вывода. Это также подразумевает большие вычислительные затраты, необходимость в оборудовании с графическим процессором и много времени на обучение. Поэтому в последнее время появилась новая тенденция – созда-

вать более легкие модели, с меньшим количеством параметров, ориентируясь скорее на качество обучающего набора данных.

Одной из последних моделей этого нового тренда является Falcon LLM, модель с открытым исходным кодом, разработанная Институтом технологических инноваций Абу-Даби. В настоящее время данная языковая модель занимает 1-е место в мире по результатам последней независимой проверки моделей искусственного интеллекта с открытым исходным кодом.

Falcon LLM прошел обучение на 1 триллионе токенов и имеет 40 миллиардов параметров. Таким образом, может возникнуть вопрос: как модель со “всего” 40 миллиардами параметров может работать так хорошо? Фактически, ответ заключается в качестве набора данных.

Falcon был разработан с использованием специализированных инструментов и включает в себя уникальный конвейер данных, способный извлекать ценный контент из веб-данных. Конвейер был разработан для извлечения высококачественного контента с использованием обширных методов фильтрации и дедупликации. Результирующий набор данных под названием RefinedWeb был выпущен компанией-разработчиком по лицензии Apache-2.0 [1].

Кроме того, архитектура Falcon была тщательно отлажена для обеспечения оптимальной производительности. Сочетая превосходное качество данных с этими оптимизациями, Falcon достигает замечательной производительности при использовании около 75% бюджета обучающих вычислений GPT-3. Более того, для вывода требуется лишь пятая часть вычислительных ресурсов.

Falcon LLM – это модель, работающая только с декодером (рис.2).

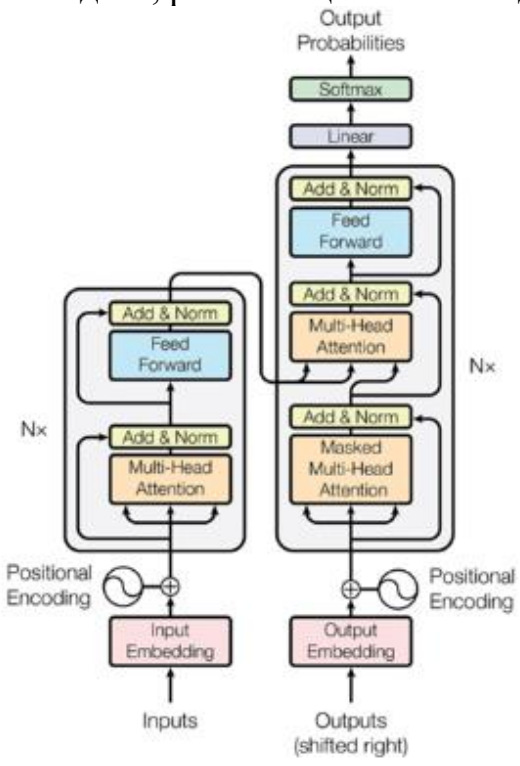

Рисунок 2 – Архитектура кодировщика-декодера

Архитектура кодировщика-декодера (Encoder-Decoder architecture) является оригинальной архитектурой transformer, представленной в документе [2]. Задача кодировщика (encoder) – представить входные данные в пространстве меньших измерений, задача декодера (decoder) – преобразовать обратно в исходный формат данные меньших измерений, предоставленные кодировщиком.

Первоначальная архитектура transformer состояла из обоих компонентов – кодировщика и декодера, однако в последние годы лаборатории, занимающиеся разработкой генеративного искусственного интеллекта, перешли к новой архитектуре, основанной только на фреймворке с декодером (например, GPT-3 OpenAI выполнен по архитектуре, основанной только на декодере). Ключевое различие между архитектурой только с декодером и архитектурой кодер-декодер заключается в отсутствии отдельного кодировщика, отвечающего за обобщение входной информации. Вместо этого в архитектуре, основанной только на декодере, скрытое состояние декодера неявно кодирует соответствующую информацию и постоянно обновляется на каждом этапе процесса генерации.

Поскольку Falcon LLM – это модель с открытым исходным кодом, то с ней можно работать из интерфейса, предоставленного в [3]. Кроме того, модель можно загрузить и использовать в программах, написанных на языке Python, для этого необходимо использовать библиотеки transformers (модули AutoTokenizer, AutoModelForCausalLM) и torch.

В зависимости от возможностей компьютерного оборудования можно выбрать модель с параметрами 40b или 7b. При этом следует учесть, что версия модели 7b обучается только английскому и французскому языкам.

Языковые модели генеративного искусственного интеллекта с открытым исходным кодом чрезвычайно мощны, и за последние несколько лет наблюдается экспоненциальный рост числа их параметров. Тем не менее, разработчики быстро приближаются к пределу, который соответствует необходимой вычислительной мощности. Поэтому крайне важно изучать новые способы сделать LLM менее “большими”, но более точными, чего добилась компания-разработчик Falcon LLM за счет того, что основное внимание уделила качеству набора обучающих выборок, что существенно повлияло на производительность модели.

Список литературы:

1. RefinedWeb. Режим доступа –https://huggingface.co/datasets/tiiuae/falcon-refinedweb.
2. The Encoder-Decoder architecture was the original transformer architecture introduced in the Attention Is All You Need (https://arxiv.org/abs/1706.03762) paper in 2017.
3. https://huggingface.co/tiiuae/falcon-rw-1b