УДК 004.8

РАЗРАБОТКА МОДЕЛИ ПРОГНОЗИРОВАНИЯ МИНИМАЛЬНО НЕОБХОДИМЫХ РЕСУРСОВ ДЛЯ РАБОТЫ ИНФЕРЕНС МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Хурматов А.М. (ИТМО)

Научный руководитель – кандидат технических наук, ст. преподаватель Ходненко И.В. (ИТМО)

Введение. Оценка времени работы нейронных сетей имеет ряд важных применений и может быть критически важной для исследователей, инженеров и практиков в области машинного обучения [1]. Знание времени работы модели позволяет эффективно распределять и оптимизировать ресурсы, такие как вычислительные мощности, для минимизации времени предсказания и улучшения общей производительности [2]. Оценки времени предсказания помогают планировать и управлять проектами в области исследований и разработок, учитывая, сколько времени потребуется для предсказания моделей и достижения поставленных целей. Предварительная оценка времени предсказания может выявить потенциальные проблемы с использованными моделями или наборами данных, помогая избежать долгих и неэффективных экспериментов, а также позволяет оценить эффективность различных методов обучения, оптимизации и архитектур [3], [4]. Для удобства использования алгоритм оценки времени возможно встроить в облачный сервер, который будет способен оценивать стоимость предсказания для моделей ИИ на различных серверах компаний, что позволит выбрать оптимальный сервис для аренды вычислительных ресурсов.

Основная часть. В ходе предварительного анализа исследований в нескольких областях и доступных реализаций методов предсказания времени работы моделей искусственного интеллекта был сформулирован подход и параметры, которые позволят реализовать модель предсказания времени работы моделей. Работа моделей разделялась на две задачи – обучение моделей ИИ и предсказание. Первым этапом является сбор и подготовка данных для обучения модели. Исходные данные представляли собой строки с гиперпараметрами каждой из варьируемых моделей, такие как batch size, learning rate, входной набор данных и число параметров модели, а также параметрами аппаратного обеспечения, на котором производилось обучение или предсказание. В качестве будущей целевой переменной замерялось время, которое потребовалось для обучения или предсказания модели на конечном устройстве при наборе заданных параметров. Вторым этапом является непосредственное обучение модели для предсказания времени работы. Для этого использовались и сравнивались три модели регрессии – Linear Regression, XGB Regressor, KNN Regression, метрикой для оценки моделей выступала MAE, как наиболее наглядная для данной задачи. Для обучения моделей ИИ лучший результат показала KNN Regression, а для предсказания моделей – XGB Regressor. Также после обучения моделей предсказания времени были выявлены параметры, наиболее сильно влияющие на время работы обучения или предсказания. Для тестирования нашего подхода мы обучили модель на одном наборе данных с аппаратным обеспечением и входными параметрами и предсказывали на другом наборе, неизвестном модели. По итогу проведенной работы алгоритм способен предсказывать время, которое потребуется для обучения или предсказания модели на том или ином конечном устройстве.

Выводы. В рамках работы были собраны данные, реализован алгоритм для предсказания времени работы моделей ИИ и были определены важные параметры для решения задачи обучения или предсказания. Основной целью исследования была разработка метода предсказания времени работы моделей ИИ на основе информации о данных, параметрах моделей и аппаратного обеспечения. Обученные на наборе параметров простые модели регрессии показывают результаты с метрикой MAE 0.3 секунды для задачи предсказания и MAE 55 секунд для задачи обучения. Были сравнены модели для предсказания времени

обучения. Модель, которая показала наилучший результат была протестирована на неизвестной для модели конфигурации.

Дальнейшие исследования могут быть направлены на оптимизацию алгоритма сбора данных, разработку cloud сервиса по предсказанию времени работы и интегрирование алгоритма оценки времени работы в многопоточном режиме.

Список использованных источников:

1. D. Justus, J. Brennan, S. Bonner and A. S. McGough, "Predicting the Computational Cost of Deep Learning Models," 2018 IEEE International Conference on Big Data (Big Data), Seattle, WA, USA, 2018, pp. 3873-3882, doi: 10.1109/BigData.2018.8622396.
2. Zancato, Luca, Alessandro Achille, Avinash Ravichandran, Rahul Bhotika and Stefano Soatto. “Predicting Training Time Without Training.” ArXiv abs/2008.12478 (2020): n. pag.
3. Dube, Parijat, Tonghoon Suk and Chen Wang. “AI Gauge: Runtime Estimation for Deep Learning in the Cloud.” 2019 31st International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD) (2019): 160-167.
4. Y. Gao, X. Gu, H. Zhang, H. Lin and M. Yang, "Runtime Performance Prediction for Deep Learning Models with Graph Neural Network," 2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP), Melbourne, Australia, 2023, pp. 368-380, doi: 10.1109/ICSE-SEIP58684.2023.00039.

Автор _____ Хурматов А.М.

Научный руководитель _____ Ходненко И.В.