

МОНИТОРИНГ И ОПТИМИЗАЦИЯ ПРОИЗВОДИТЕЛЬНОСТИ ИИ-СИСТЕМ В РЕАЛЬНОМ ВРЕМЕНИ

Шешукова А.В.

МИРЭА – Российский технологический университет, Москва

Ключевые слова: искусственный интеллект, мониторинг производительности, оптимизация ресурсов, машинное обучение, real-time системы, предиктивная аналитика.

Аннотация. В статье рассматривается проблема оптимизации производительности систем искусственного интеллекта в реальном времени. Предложена гибридная архитектура мониторинга, сочетающая традиционные методы с машинным обучением. Проведено экспериментальное исследование, показавшее повышение производительности на 25-30% и снижение затрат на инфраструктуру на 25%. Результаты могут применяться для масштабирования существующих ИИ-систем.

MONITORING AND OPTIMIZING THE PERFORMANCE OF AI SYSTEMS IN REAL TIME

Sheshukova A.V.

MIREA – Russian Technological University, Moscow

Keywords: artificial intelligence, performance monitoring, resource optimization, machine learning, real-time systems, predictive analytics.

Abstract. The article discusses the problem of optimizing the performance of artificial intelligence systems in real time. A hybrid monitoring architecture is proposed that combines traditional methods with machine learning. An experimental study was conducted that showed a 25-30% increase in productivity and a 25% reduction in infrastructure costs. The results can be applied to scale existing AI systems.

Введение

В условиях стремительного развития технологий искусственного интеллекта (ИИ) и их внедрения в различные сферы жизни, мониторинг и оптимизация производительности таких систем становятся важнейшими аспектами их эффективного использования. Современные ИИ-системы находят применение в области здравоохранения, финансов, промышленности и многих других, где от их производительности зависят ключевые показатели эффективности. Однако высокая сложность и динамичность этих систем требуют применения специальных подходов для их мониторинга и оптимизации в реальном времени. Это позволяет своевременно выявлять и устранять проблемы, обеспечивая надежность и точность работы ИИ.

Целью данного исследования является изучение методов мониторинга и оптимизации производительности ИИ-систем в реальном времени. Для достижения этой цели ставятся следующие задачи: анализ существующих подходов к мониторингу, исследование технологий сбора и анализа данных, а также разработка и оценка стратегий оптимизации. Работа направлена на выявление наиболее эффективных методов и их адаптацию для применения в

различных отраслях, что позволит повысить производительность и надежность ИИ-систем.

Теоретические основы мониторинга ИИ-систем

Современные системы искусственного интеллекта характеризуются высокой вариативностью вычислительных нагрузок. Основные факторы, влияющие на распределение ресурсов: сложность нейронных сетей, объем входных данных, требования к времени обработки, тип выполняемых операций (инференс/обучение).

Традиционные методы мониторинга имеют ряд недостатков для динамических ИИ-систем: высокая задержка в принятии решений, статичность правил распределения ресурсов, отсутствие предиктивной аналитики.

Использование адаптивных систем для повышения эффективности

Адаптивные системы представляют собой динамические структуры, способные автоматически изменять свои параметры для достижения оптимальной производительности в условиях изменяющейся среды. Основой их работы является использование методов машинного обучения, которые анализируют входные данные и принимают решения в реальном времени. Это достигается за счет создания моделей, учитывающих текущие условия и предсказывающих наиболее эффективные действия. Например, такие системы могут адаптироваться к изменениям нагрузки на серверы, перераспределяя ресурсы для обеспечения стабильной работы. При этом методы и инструментальные средства для решения задач дискретной оптимизации на распределенных системах подчеркивают важность применения параллельных вычислений, что способствует повышению производительности ИИ-систем [1]. Адаптивные системы, таким образом, играют ключевую роль в повышении эффективности работы ИИ-систем, обеспечивая их гибкость и устойчивость в условиях неопределенности.

Применение адаптивных систем в реальном времени демонстрирует их значительный потенциал в различных отраслях. Например, алгоритм управления энергопотреблением в дата-центрах компании Google использует машинное обучение для анализа данных о температуре, нагрузке на оборудование и других параметрах, что позволяет оптимизировать работу систем охлаждения. В результате удалось снизить затраты на электроэнергию на 40%, что не только уменьшило расходы, но и сократило углеродный след.

Предлагаемая система

Предложенная архитектура системы включает следующие модули: модуль сбора метрик с микросекундной точностью, модуль предиктивной аналитики, модуль принятия решений, модуль автоматического масштабирования.

Методы оптимизации, используемые в системе – это динамическое масштабирование ресурсов, квантизация моделей, пакетная обработка запросов, кэширование результатов, оптимизация графа вычислений.

Ключевые отслеживаемые параметры, с помощью которых проверяется выдвигаемая гипотеза: загрузка CPU/GPU, потребление памяти, сетевая активность, время отклика, точность модели.

Алгоритм оптимизации ресурсов состоит из следующих этапов: сбор и агрегация метрик, прогнозирование нагрузки, оценка оптимального распределения ресурсов, автоматическое масштабирование.

Программные средства для реализации системы: TensorFlow/PyTorch для инференса, Prometheus/Grafana для мониторинга, Kubernetes для оркестрации, Apache Kafka для обработки событий.

Тестирование системы проводилось на следующих сценариях: стабильная нагрузка, резкие скачки нагрузки, постепенное нарастание нагрузки, смешанная нагрузка. Исходные данные для тестирования ИИ-модели: распознавания человека на картинке (Yola), модель кластеризации, модель классификации, модель распознавания какое животное на картинке. Во всех случаях использовались уже готовые нейронные сети и открытые датасеты. Производительная среда с нагрузкой 1000-5000 запросов в секунду.

Результаты экспериментов: снижение времени отклика на 20-25%, повышение пропускной способности на 30%, уменьшение затрат на инфраструктуру на 25%, стабильность работы при пиковых нагрузках.

Количественные показатели: среднее время отклика 150-250 мс, процент успешных оптимизаций 85%, коэффициент использования ресурсов 80%. Качественные результаты исследования – стабильная работа системы при пиковых нагрузках, автоматическое восстановление после сбоев и эффективное распределение ресурсов.

При проведении эксперимента были выявлены следующие проблемы: задержки при масштабировании, проблемы с кэшированием при обновлении моделей, ограничение сетевой инфраструктуры. Конструктивное решение состоит в внедрение предиктивного масштабирования, оптимизации кэширования, балансировки нагрузки.

Заключение

Проведенное исследование демонстрирует эффективность предложенного подхода к мониторингу и оптимизации ИИ-систем. Основные выводы:

1. Гибридный подход, сочетающий традиционные методы мониторинга с машинным обучением, обеспечивает оптимальное управление ресурсами.
2. Внедрение разработанной системы позволяет повысить производительность на 25-30%.
3. Предложенная архитектура обеспечивает микросекундную точность мониторинга.

Практическая значимость работы заключается в возможности масштабируемого внедрения разработанных решений в существующие ИИ-системы, что способствует повышению их эффективности и снижению операционных затрат.

Научная значимость работы заключается в теоретическом и методологическом аспектах. Теоретический вклад заключается в разработке комплексного подхода к мониторингу ИИ-систем в реальном времени. Методологическое значение состоит в создании алгоритма предиктивного масштабирования ресурсов и внедрение инновационных методов автоматической оптимизации.

Список литературы

1. Афанасьев А.П., Волошинов В.В., Посыпкин М.А., Сигал И.Х., Хуторной Д.А. Программный комплекс для решения задач дискретной оптимизации на распределенных вычислительных системах // Труды ИСА РАН. – 2006. – Т. 25. – С.5-17.
2. Перевозникова Н.В., Мызникова М.А «Оценка эффективности стратегии цифровой трансформации» [Электронный ресурс]. – Режим доступа: <http://repo.donnu.ru:8080/jspui/bitstream/123456789/4807/1/3556.pdf#page=413>.
3. Секисов А.Н., Коженко Н.В., Папоян А.А., Кристя Н.Г., Прозорова А.С. «Основные тенденции и направления применения искусственного интеллекта в строительном секторе национальноного хозяйства: организационные и экономические аспекты» // Экономика: вчера, сегодня, завтра – 2023 – Т. 13, №10А. – С. 357-366. – DOI: 10.34670/AR.2023.20.36.039.

Сведения об авторе:

Шешукова Анна Викторовна – аспирант.