ФЕДЕРАЛЬНОЕ АГЕНТСТВО

ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

ПНСТ 945— 2024

Искусственный интеллект

ТЕХНИЧЕСКАЯ СТРУКТУРА ДЛЯ РАЗДЕЛЕНИЯ И СОВМЕСТНОГО ИСПОЛНЕНИЯ МОДЕЛИ ГЛУБОКОЙ НЕЙРОННОЙ СЕТИ

[ITU-T F.748.20 (2022), NEQ]

Издание официальное

Москва
Российский институт стандартизации
2024

Предисловие

1 РАЗРАБОТАН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова» (ФГБОУ ВО МГУ имени М.В. Ломоносова) и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ООО «ИРИО»)

2 ВНЕСЕН Техническим комитетом по стандартизации ТК 164 «Искусственный интеллект»

3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 3 октября 2024 г. № 54-пнст

4 Настоящий стандарт разработан с учетом основных нормативных положений международного документа ITU-T F.748.20 (12/2022) «Техническая структура для разделения и совместного исполнения модели глубокой нейронной сети» (ITU-T F.748.20 (12/2022) «Technical framework for deep neural network model partition and collaborative execution», NEQ)

Правила применения настоящего стандарта и проведения его мониторинга установлены в ГОСТ Р 1.16—2011 (разделы 5 и 6).

Федеральное агентство по техническому регулированию и метрологии собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее чем за 4 мес до истечения срока его действия разработчику настоящего стандарта по адресу: 119991 Москва, Ленинские горы, д. 1, и в Федеральное агентство по техническому регулированию и метрологии по адресу: 123112 Москва, Пресненская набережная, д. 10, стр. 2.

В случае отмены настоящего стандарта соответствующая информация будет опубликована в ежемесячном информационном указателе «Национальные стандарты» и также будет размещена на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет (www.rst.gov.ru)

Содержание

Введение

Процесс логического вывода для модели глубокой нейронной сети, как правило, требует больших объемов вычислительных ресурсов и памяти. В связи с этим самостоятельное исполнение модели глубокой нейронной сети оконечными устройствами представляется сложной задачей. Эффективным способом реализации совместного исполнения глубокой нейронной сети посредством оконечных и периферийных устройств является разделение модели глубокой нейронной сети, которое может снизить задержку исполнения и одновременно улучшить использование ресурсов. Целью настоящего стандарта является определение технической структуры разделения и совместного исполнения модели глубокой нейронной сети. Во-первых, необходимо заранее спрогнозировать общую задержку логического вывода, учитывая текущее состояние системы в соответствии с различными стратегиями разделения глубокой нейронной сети. Затем, опираясь на вычислительные возможности оборудования, состояние сети и свойства модели глубокой нейронной сети, необходимо выбрать подходящие места для разделения и стратегию совместного исполнения. На последнем этапе реализуется совместное исполнение модели и оптимизируется распределение ресурсов.

ПРЕДВАРИТЕЛЬНЫЙ НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

Искусственный интеллект

ТЕХНИЧЕСКАЯ СТРУКТУРА ДЛЯ РАЗДЕЛЕНИЯ И СОВМЕСТНОГО ИСПОЛНЕНИЯ МОДЕЛИ ГЛУБОКОЙ НЕЙРОННОЙ СЕТИ

Artificial intelligence. Technical framework for deep neural network model partition and collaborative execution

Срок действия — с 2025—01—01
до 2028—01—01

1 Область применения

Настоящий стандарт направлен на определение технической структуры разделения и совместного исполнения модели глубокой нейронной сети (ГНС), включая создание модели прогнозирования задержки, разработку стратегии разделения модели ГНС, оптимизацию распределения ресурсов и совместное исполнение модели для соответствия требованиям логического вывода модели ГНС и обеспечения эффективного использования ресурсов между оконечными устройствами и периферийным устройством.

Область применения настоящего стандарта включает:

- подходы к разделению и совместному исполнению модели ГНС,
- создание модели прогнозирования задержки для разделения модели ГНС,
- разработку стратегии разделения модели ГНС,
- совместное исполнение модели ГНС после ее разделения.

2 Нормативные ссылки

В настоящем стандарте использованы нормативные ссылки на следующие стандарты:

ГОСТ Р 71476 Искусственный интеллект. Концепции и терминология искусственного интеллекта

ГОСТ Р 34.13—2015 Информационная технология. Криптографическая защита информации. Режимы работы блочных шифров

ПНСТ 845—2023 Искусственный интеллект. Техническая структура федеративной системы машинного обучения

Примечание — При использовании настоящим стандартом целесообразно проверить действие ссылочных стандартов в информационной системе общего пользования — на официальном сайте Федерального агентства по техническому регулированию и метрологии в сети Интернет или по ежегодному информационному указателю «Национальные стандарты», который опубликован по состоянию на 1 января текущего года, и по выпускам ежемесячного информационного указателя «Национальные стандарты» за текущий год. Если заменен ссылочный стандарт, на который дана недатированная ссылка, то рекомендуется использовать действующую версию этого стандарта с учетом всех внесенных в данную версию изменений. Если заменен ссылочный стандарт, на который дана датированная ссылка, то рекомендуется использовать версию этого стандарта с указанным выше годом утверждения (принятия). Если после утверждения настоящего стандарта в ссылочный стандарт, на который дана датированная ссылка, внесено изменение, затрагивающее положение, на которое дана ссылка, то это положение рекомендуется применять без учета данного изменения. Если ссылочный стандарт отменен без замены, то положение, в котором дана ссылка на него, рекомендуется применять в части, не затрагивающей эту ссылку.

Издание официальное

3 Термины и определения

В настоящем стандарте применены термины по ГОСТ Р 71476, а также следующие термины с соответствующими определениями:

3.1 **логический вывод модели глубокой нейронной сети** (deep neural network model inference): Процесс исполнения всех слоев глубокой нейронной сети последовательно, и после завершения последнего слоя глубокой нейронной сети получение результата логического вывода.

3.2

модель глубокого обучения (deep learning model): Алгоритм глубокого обучения для решения конкретной задачи, обычно содержащий информацию о структуре вычислительного графа и информацию о параметрах, используемых для представления алгоритма глубокого обучения.

[ПНСТ 844—2023, пункт 2.3]

3.3 **разделение модели глубокой нейронной сети** (deep neural network model partition): Процесс разделения на две или более части модели глубокой нейронной сети в соответствии со стратегиями разделения, и для каждой части выбирается свое место для исполнения, например, оконечное/периферийное устройство.

4 Сокращения

В настоящем стандарте применены следующие сокращения:

ГНС — глубокая нейронная сеть;
ПДн — персональные данные;
ADA — алгоритм (adaboost);
ANN — искусственная нейронная сеть (artificial neural network);
DT — дерево решений (decision tree);
GBRT — дерево регрессии с повышением градиента (gradient boosting regression tree);
KNN — алгоритм k-ближайших соседей (k-nearest neighbour);
KRR — регрессия гребня ядра (kernel ridge regression);
LR — линейная регрессия (linear regression);
MAE — средняя абсолютная ошибка (mean absolute error);
MAPE — средняя абсолютная ошибка в процентах (mean absolute percentage error);
RANSAC — регрессия консенсуса по случайной выборке (random sample consensus regression);
RF — случайный лес (random forest);
SVM — метод опорных векторов (support vector machine).

5 Соглашения по терминологии

В настоящем стандарте:

- ключевые слова «требуется, чтобы» означают требование, которое должно строго соблюдаться и отклонение от которого не допускается, если будет сделано заявление о соответствии этому документу;
- ключевое слово «рекомендуется» означает требование, которое рекомендуется, но не является абсолютно необходимым. Таким образом, это требование не является обязательным для заявления о соответствии настоящему документу.

6 Обзор разделения и совместного исполнения модели глубокой нейронной сети

Общая структура разделения и совместного исполнения (посредством оконечных/периферийных устройств) модели ГНС показана на рисунке 1. Общая структура состоит из периферийного устройства и нескольких оконечных устройств, которые включают в себя модуль прогнозирования задержки вычислений, модуль мониторинга сети, модуль процессора модели, модуль разделения ГНС, модуль обработки запросов и модуль инструмента развертывания.

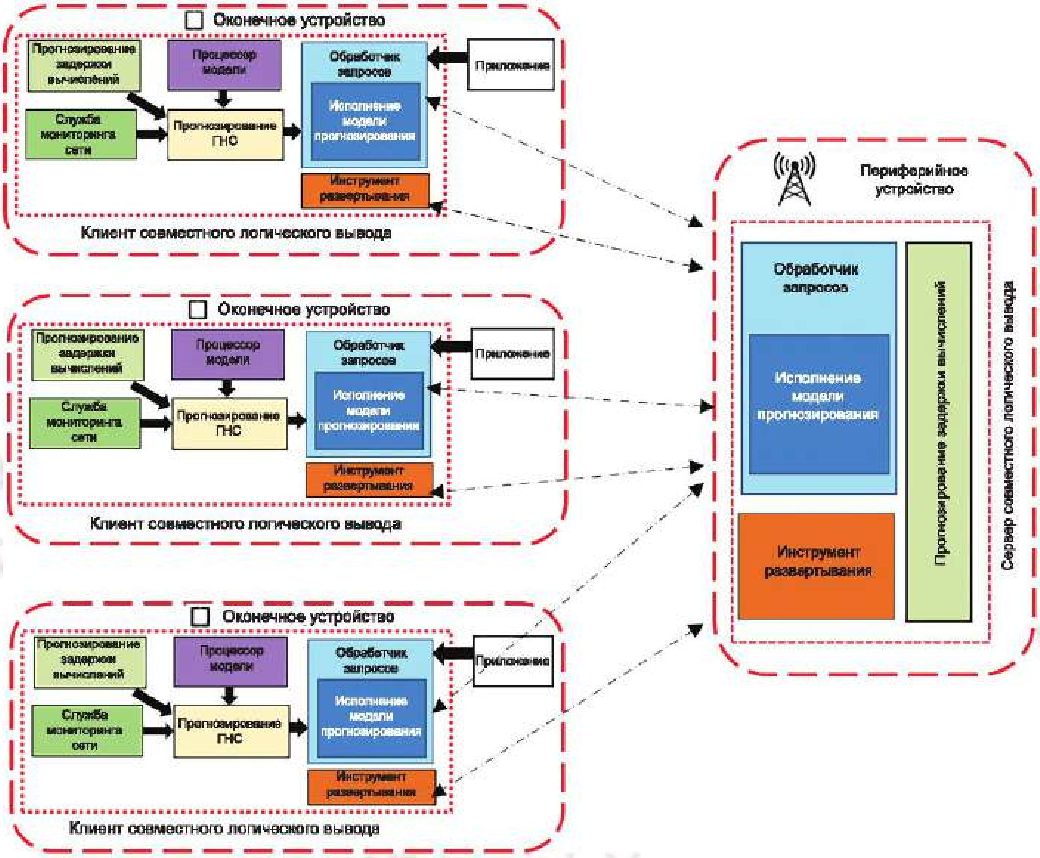

Рисунок 1 — Общая структура разделения и совместного исполнения модели ГНС

Модуль инструмента развертывания используется для сбора данных о задержке обучения в различных конфигурациях системы, для управления использованием центрального процессора и памяти оконечных устройств и периферийного устройства.

Модуль процессора модели используется для анализа и извлечения топологии модели ГНС, статических характеристик слоев ГНС, размера выходных данных ГНС и т. д.

Модуль прогнозирования задержки вычислений используется для прогнозирования задержки вычислений каждого слоя ГНС. Данный модуль выполняется с использованием метрических данных, собранных модулем инструмента развертывания, который учитывает многочисленные статические и динамические характеристики системы.

Модуль мониторинга сети используется для мониторинга пропускной способности сети при загрузке/скачивании в режиме реального времени и для оценки задержки передачи на слоях ГНС.

Модуль разделения ГНС используется для определения подходящих мест разделения для совместного исполнения (посредством оконечных/периферийных устройств) в соответствии с информацией, полученной от модуля процессора модели, модуля прогнозирования задержки вычислений и модуля мониторинга сети.

Модуль обработки запросов используется для обработки в режиме реального времени запросов от оконечных устройств на логический вывод модели ГНС. Данный модуль контролирует оконечные устройства и периферийное устройство при совместном исполнении логического вывода ГНС в соответствии со схемой разделения.

7 Создание модели прогнозирования задержки для разделения модели глубокой нейронной сети

7.1 Мониторинг состояния системы и настройка параметров

В модель ГНС входят несколько слоев разного типа, таких как сверточный слой, полносвязный слой, слой пулинга, слой активации. Каждый тип слоя нейронной сети имеет разные конфигурации параметров и необходим для создания разных моделей прогнозирования задержки. Сверточный слой и полносвязный слой являются наиболее распространенными. Основные параметры сверточного слоя включают размер входных данных, размер ядра свертки, размер канала, шаг свертки и дополнение (см. ГОСТ Р 34.13—2015, пункт 2.1.6). Основными параметрами полносвязного слоя являются размеры входных и выходных данных.

Помимо конфигурации слоя состояние системы также оказывает существенное влияние на задержку вычислений для данного слоя. Необходимо отслеживать динамическое состояние системы, включая загрузку центрального процессора, объем свободной памяти, использование памяти, пропускную способность сети, топологию сети и так далее. Для профилирования модели прогнозирования задержки необходимо точно и своевременно мониторить перечисленные выше динамические состояния системы.

7.2 Моделирование и анализ прогнозирования задержки

Для каждого типа слоев нейронной сети необходимо создать модель прогнозирования задержки вычислений. Входные данные модели прогнозирования задержки включают конфигурацию параметров слоя и состояние системы. Выходом модели предсказания задержки является значение задержки вычислений для слоя с учетом текущей конфигурации параметров слоя и состояния системы.

Существует много моделей прогнозирования задержки, например, LR (линейная регрессия), RANSAC (регрессия консенсуса по случайной выборке), KRR (регрессия гребня ядра), KNN (алгоритм k-ближайших соседей), DT (дерево решений), SVM (метод опорных векторов), RF (случайный лес), ADA (алгоритм), GBRT (дерево регрессии с повышением градиента), ANN (искусственная нейронная сеть).

Для выбора модели прогнозирования задержки необходимо сравнить несколько моделей и определить наиболее подходящую. Показатели оценки включают MAE (средняя абсолютная ошибка) и MAPE (средняя абсолютная ошибка в процентах).

8 Разработка стратегии разделения модели глубокой нейронной сети

8.1 Анализ задержки исполнения модели при различных стратегиях разделения для разнообразных состояний системы

Общая задержка исполнения на каждом оконечном устройстве состоит из трех частей: задержка локальных вычислений на оконечном устройстве, задержка связи между оконечным и периферийным устройствами и задержка вычислений на периферийном устройстве. На общую задержку исполнения могут повлиять различные стратегии разделения модели ГНС.

В частности, на задержку локальных вычислений оказывают влияние вычислительные возможности оконечного устройства и количество слоев ГНС, исполняемых локально. Задержка связи определяется исходя из пропускной способности выделенных коммуникационных ресурсов и объемов промежуточных данных. Задержка периферийных вычислений зависит от величины выделенных ресурсов периферийных вычислений и количества оставшихся слоев ГНС.

Когда состояние системы, а именно пропускная способность выделенных коммуникационных и вычислительных ресурсов, изменяется, требуется соответствующим образом изменить и стратегии разделения, направленные на достижение компромисса между задержкой вычислений и задержкой связи.

8.2 Подтверждение позиции разделения и распределение ресурсов

Анализ задержки исполнения модели при различных стратегиях разделения для разнообразных состояний системы показывает необходимость одновременно оптимизировать стратегию разделения ГНС и распределять ресурсы, чтобы минимизировать задержку или потребление энергии. Стратегия разделения модели ГНС предполагает ее разделение на две и более части путем определения соответ-

ствующих мест разделения и выбора места исполнения (т. е. оконечного/периферийного устройства) для каждого слоя ГНС. Модели ГНС можно разбить на две категории: ГНС с цепной структурой (chain DNN) и ГНС с нецепной структурой (non-chain DNN). Следовательно, необходимо разработать различные стратегии разбиения для двух категорий моделей ГНС. Распределение ресурсов в основном включает в себя распределение ресурсов связи и вычислительных ресурсов периферийного устройства.

Место разделения определяется стратегией разделения. Структура на рисунке 1 демонстрирует взаимодействие многочисленных оконечных устройств с периферийным устройством, имеющим ограниченные ресурсы. Для каждого оконечного устройства нужно принять решение о разделении в соответствии с вычислительными возможностями оконечных устройств и выделенными периферийными ресурсами. Если вычислительных возможностей оконечного устройства достаточно, то его модель ГНС не разделяется и не занимает ресурсы периферийного устройства, при этом вся модель ГНС полностью исполняется на оконечном устройстве. Если же вычислительных возможностей оконечного устройства недостаточно, то его модель ГНС будет разделена и совместно исполняться оконечным и периферийным устройствами.

Более того, стратегию разделения необходимо осуществлять одновременно с распределением ресурсов между оконечными устройствами. В частности, периферийное устройство должно эффективно выделять свои вычислительные ресурсы каждому оконечному устройству для исполнения выгруженных с оконечных устройств слоев ГНС, а также выделять свои коммуникационные ресурсы каждому оконечному устройству для передачи промежуточных данных.

9 Совместное исполнение модели глубокой нейронной сети после разделения

9.1 Развертывание модели ГНС на основе стратегии разделения

После выбора стратегии разделения модели ГНС определяется место исполнения каждого слоя ГНС — либо на периферийном, либо на оконечном устройстве. Из-за разнообразия и вариаций моделей ГНС представляется нецелесообразным заранее развертывать все модели ГНС на периферийном устройстве.

В зависимости от выбранной стратегии разделения выделяют два варианта развертывания моделей ГНС по запросу. Первый вариант заключается в развертывании всей модели ГНС на периферийном устройстве. Второй вариант заключается в развертывании необходимых слоев ГНС, которые для исполнения распределены на периферийное устройство. Исполнение модели ГНС возможно только после развертывания модели ГНС или необходимых слоев ГНС.

9.2 Процесс логического вывода при совместном исполнении модели ГНС

В соответствии со стратегией разделения каждый слой ГНС последовательно исполняется либо на периферийном устройстве, либо на оконечном устройстве (см. ПНСТ 845—2023). Каждый слой ГНС получает выходные данные своих предшествующих слоев в качестве собственных входных данных, исполняет ГНС и передает свои выходные данные своим последующим слоям. Если предшествующий слой и его последующий слой исполняются на разных устройствах (например, предшествующий слой исполняется на периферийном устройстве, а его последующий слой исполняется на оконечном устройстве, или предшествующий слой исполняется на оконечном устройстве, а его последующий слой исполняется на периферийном устройстве), то выходные данные предшествующего слоя должны передаваться по каналам связи на его последующий слой. Если предшествующий слой и его последующий слой исполняются на одних и тех же устройствах (например, предшествующий слой и его последующий слой исполняются на одном оконечном устройстве или на периферийном устройстве), то выходные данные предшествующего слоя будут переданы напрямую в его последующий слой. Последующий слой может быть исполнен лишь после того, как он получит выходные данные своего предшествующего слоя. Процесс логического вывода завершается, когда все слои ГНС исполнены.

Приложение А
(справочное)

Типичные сценарии для разделения и совместного исполнения модели глубокой нейронной сети

А.1 Умный дом

Многие приложения для умного дома, в частности, безопасность дома и видеонаблюдение, являются частью нашей повседневной жизни. Данные приложения должны исполнять модели ГНС, такие как компьютерное зрение, распознавание видео или машинный перевод. Однако большинство домашних устройств, включая мобильные телефоны, камеры и различные датчики, не способны самостоятельно исполнять модели ГНС, что приводит к большой задержке исполнения и ухудшению пользовательского опыта. В этом сценарии следует применять совместное исполнение ГНС оконечными и периферийными устройствами. Периферийные серверы и базовые станции являются периферийными устройствами и взаимодействуют с домашними устройствами для исполнения моделей ГНС, что позволяет снизить задержку исполнения и одновременно защитить ПДн.

А.2 Промышленное производство

Спрос на интеллектуальные технологии в промышленном производстве стремительно растет. В действительности интеллект является ключом к цифровой трансформации промышленного производства. Для промышленного процесса необходим анализ данных и наличие интеллектуальных возможностей принятия решений, которые зависят от различных интеллектуальных приложений и соответствующих моделей ГНС. Например, интеллектуальный мониторинг, распознавание изображений и другие приложения применяются для мониторинга безопасности на производстве и устранения рисков для безопасности. Для обеспечения безопасности и своевременности интеллектуального промышленного производства необходимо реализовать совместное исполнение ГНС посредством ее разделения.

А.3 Интеллектуальная транспортная система

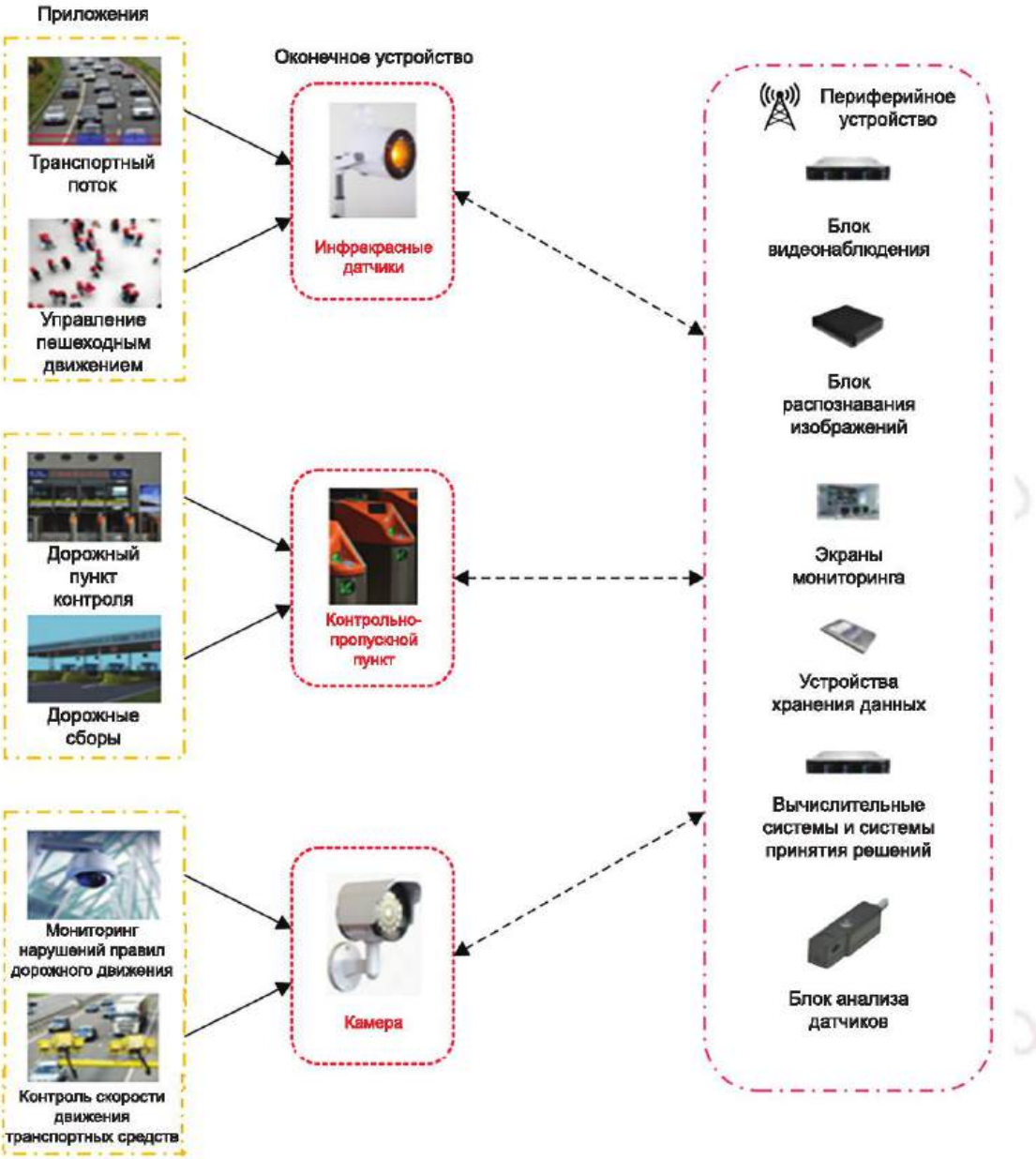

Рисунок А.1 — Разделение модели ГНС и совместное исполнение в интеллектуальном транспортном приложении

Многие приложения, использующие искусственный интеллект для управления дорожным движением, такие как мониторинг транспортных потоков, управление пешеходным движением, контрольно-пропускные пункты, дорожные сборы, мониторинг нарушений и скорости движения, широко распространены в нашей жизни. Эти приложения должны исполнять модели ГНС, реализующие распознавание изображений, семантическую сегментацию (например, распознавание лиц), распознавание номерных знаков, мониторинг поведения и управление транспортными потоками.

Для функционирования транспортных приложений, использующих искусственный интеллект, необходимы различные типы интеллектуальных оконечных устройств, таких как инфракрасные датчики, контрольно-пропускные устройства и интеллектуальные камеры, которые обладают некоторой вычислительной мощностью. Однако текущий метод исполнения зависит только от периферийных устройств, которые должны самостоятельно обраба-

тывать большие объемы данных, что оказывает существенную нагрузку на оборудование. Традиционные ГНС не в полной мере обеспечивают вычислительные и аналитические возможности оконечных устройств. Поэтому система обладает относительно низкой эффективностью совместной работы и высокой задержкой исполнения.

Для того чтобы в полной мере использовать вычислительные возможности каждого устройства, применяются стратегии разделения модели ГНС и совместного исполнения с распределением некоторых слоев ГНС от периферийных устройств к целевому оконечному устройству. Оконечные и периферийные устройства совместно выполняют задачи вычислений и анализа, что повышает эффективность исполнения. Реализация соответствующих стратегий разделения для совместной работы оконечных и периферийных устройств в интеллектуальных транспортных системах (показанных на рисунке А.1) позволяет удовлетворить потребности в интеллектуальных транспортных приложениях.

Ниже перечислены некоторые типичные интеллектуальные транспортные приложения:

а) инфракрасные датчики вместе с другими датчиками и большими экранами могут отслеживать транспортные средства и потоки пешеходов;

б) контрольно-пропускной пункт во взаимодействии с системой распознавания изображений собирает и вычисляет огромное количество информации о людях и транспортных средствах для управления дорожным движением;

в) камеры и устройства хранения данных, вычислительные системы и системы принятия решений совместно отслеживают нарушения и скорость движения.

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: информационные технологии, искусственный интеллект, техническая структура, глубокая нейронная сеть

Технический редактор *И.Е. Черепкова*
Корректор *Р.А. Ментова*
Компьютерная верстка *А.Н. Золотаревой*

Сдано в набор 07.10.2024. Подписано в печать 21.10.2024. Формат 60×84⅛. Гарнитура Ариал.
Усл. печ. л. 1,40. Уч.-изд. л. 1,18.

Подготовлено на основе электронной версии, предоставленной разработчиком стандарта

Создано в единичном исполнении в ФГБУ «Институт стандартизации» для комплектования Федерального информационного фонда стандартов,
117418 Москва, Нахимовский пр-т, д. 31, к. 2.
www.gostinfo.ru info@gostinfo.ru