ФЕДЕРАЛЬНОЕ АГЕНТСТВО
ПО ТЕХНИЧЕСКОМУ РЕГУЛИРОВАНИЮ И МЕТРОЛОГИИ

П Р Е Д В А Р И Т Е Л Ь Н Ы Й
Н А Ц И О Н А Л Ь Н Ы Й
С Т А Н Д А Р Т
Р О С С И Й С К О Й
Ф Е Д Е Р А Ц И И

ПНСТ
(проект)

Искусственный интеллект

МЕТОДОЛОГИЯ ОЦЕНКИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ

(ITU F.748.12 (2021), NEQ)

Издание официальное

Настоящий проект стандарта не подлежит применению до его утверждения

Москва
Российский институт стандартизации
2023

Предисловие

1 ПОДГОТОВЛЕН Научно-образовательным центром компетенций в области цифровой экономики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В.Ломоносова» (МГУ имени М.В.Ломоносова) и Обществом с ограниченной ответственностью «Институт развития информационного общества» (ИРИО) на основе собственного перевода на русский язык англоязычной версии документа, указанного в пункте 4

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 202 г. № -ст

4 Настоящий стандарт разработан на основе международного документа ITU-T F.748.12 (2021) «Методология оценки разработки программного обеспечения для глубокого обучения» (Recommendation ITU-T F.748.12 (2021), Deep learning software framework evaluation methodology, NEQ). В настоящий стандарт включены дополнительные по отношению к международному документу ITU-T F.748.12 определения и положения из международного стандарта ISO/IEC 22989 «Информационные технологии. Искусственный интеллект. Понятия и терминология искусственного интеллекта» (ISO/IEC 22989:2022 «Information technology — Artificial intelligence — Artificial intelligence concepts and terminology»).

Сведения о соответствии ссылочных национальных стандартов международным стандартам, использованным в качестве ссылочных в примененном международном операционном документе, приведены в дополнительном приложении ДА.

5 ВВЕДЕН ВПЕРВЫЕ

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

Национальный орган Российской Федерации по стандартизации собирает сведения о практическом применении настоящего стандарта. Данные сведения, а также замечания и предложения по содержанию стандарта можно направить не позднее, чем за девять месяцев до истечения срока его действия, разработчику

настоящего стандарта по адресу: 119991, Российская Федерация, Москва, Ленинские горы, д. 1 и в национальный орган Российской Федерации по стандартизации по адресу: 123112, город Москва, Пресненская набережная, дом 10, строение 2.

В случае отмены настоящего стандарта соответствующее уведомление будет опубликовано в ежемесячно издаваемых информационном указателе «Национальные стандарты» и журнале «Вестник технического регулирования». Уведомление будет размещено также на официальном сайте национального органа Российской Федерации по стандартизации в сети Интернет.

© ISO, 2023
© IEC, 2023
© Оформление. ФГБУ «РСТ», 2023

Настоящий стандарт не может быть полностью или частично воспроизведен, тиражирован и распространен в качестве официального издания без разрешения Федерального агентства по техническому регулированию и метрологии

Содержание

Предисловие II

Введение V

1 Область применения 1

2 Нормативные ссылки 1

3 Термины и определения 2

3.1 Термины, определенные в других документах 2

3.2 Термины, определенные в настоящем документе 2

4 Сокращения 3

5 Соглашения по терминологии 3

6 Промышленная реализация среды разработки программного обеспечения для глубокого обучения в области искусственного интеллекта 4

6.1 Архитектура приложения на основе среды разработки программного обеспечения для глубокого обучения 4

6.2 Аппаратное обеспечение 5

6.3 Компиляторы 5

6.4 Среды разработки 5

6.5 Базовые приложения 5

6.6 Отраслевые приложения 5

7 Определение требований и методов оценки среды разработки программного обеспечения для глубокого обучения 6

7.1 Обзор требований 6

7.2 Классические модели глубокого обучения 7

7.3 Конкретные показатели и методы оценки среды разработки для обучения 8

7.4 Конкретные показатели и методы оценки среды разработки для вывода 12

Приложение ДА (справочное) Сведения о соответствии ссылочных национальных стандартов ссылочным международным стандартам, использованным в качестве ссылочных в примененном международном документе 15

Введение

Среда разработки программного обеспечения для глубокого обучения предоставляет производителям простой и быстрый способ разработки собственных приложений искусственного интеллекта (ИИ). Однако разные среды разработки показывают разную производительность в различных сценариях. Настоящий стандарт, основанный на Рекомендации ITU-T F.748.12 помогает оценить среды разработки программного обеспечения для глубокого обучения, чтобы помочь производителям в полной мере использовать преимущества одних и избежать недостатков других.

В настоящий стандарт включены дополнительные по отношению к Рекомендации ITU-T F.748.12 определения и положения из международного стандарта ISO/IEC 22989 «Информационные технологии. Искусственный интеллект. Понятия и терминология искусственного интеллекта» (ISO/IEC 22989:2022 «Information technology — Artificial intelligence — Artificial intelligence concepts and terminology»). Это позволяет гармонизировать данный документ с принятыми ранее национальными стандартами и предварительными национальными стандартами в области ИИ.

Для получения доступа к данной Рекомендации наберите в адресном поле вашего веб-браузера URL-адрес http://handle.itu.int/, а затем уникальный идентификатор этой Рекомендации: http://handle.itu.int/11.1002/1000/14681

ПНСТ
(проект)

П Р Е Д В А Р И Т Е Л Ь Н Ы Й Н А Ц И О Н А Л Ь Н Ы Й С Т А Н Д А Р Т
Р О С С И Й С К О Й Ф Е Д Е Р А Ц И И

Искусственный интеллект

МЕТОДОЛОГИЯ ОЦЕНКИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ

Artificial intelligence. Deep learning software framework evaluation methodology

Дата введения –

1 Область применения

В настоящей Рекомендации представлена методология оценки среды разработки программного обеспечения для глубокого обучения. В нее рассматриваются следующие темы:

а) экосистема среды разработки программного обеспечения для глубокого обучения;

б) простота использования среды разработки программного обеспечения для глубокого обучения;

в) производительность среды разработки программного обеспечения для глубокого обучения;

г) поддерживаемая архитектура среды разработки программного обеспечения для глубокого обучения;

д) оптимизация на уровне аппаратного обеспечения среды разработки программного обеспечения для глубокого обучения;

е) безопасность и стабильность функционирования среды разработки программного обеспечения для глубокого обучения.

2 Нормативные ссылки

Указанные ниже Рекомендации МСЭ-Т и другие справочные документы содержат положения, которые путем ссылки на них в данном тексте составляют

Издание официальное

положения настоящей Рекомендации. На момент публикации указанные издания были действующими. Все Рекомендации и другие источники могут подвергаться пересмотру; поэтому всем пользователям данной Рекомендации предлагается изучить возможность применения последнего издания Рекомендаций и других справочных документов, перечисленных ниже. Перечень действующих в настоящее время Рекомендаций МСЭ-Т регулярно публикуется. Ссылка на документ в рамках данной Рекомендации не придает ему как отдельному документу статус Рекомендации.

Следующие стандарты упоминаются в тексте таким образом, что часть или все их содержание представляют собой требования настоящего стандарта. Для датированных ссылок применяется только указанное издание. Для недатированных ссылок применяется последнее издание стандарта, на который дана ссылка (включая любые поправки):

ISO/IEC 22989:2022, Информационные технологии — Искусственный интеллект — Концепции искусственного интеллекта и терминология

3 Термины и определения

3.1 Термины, определенные в других документах

Для целей настоящего документа применяются термины и определения, приведенные в ISO/IEC 22989, а также следующие.

ИСО и МЭК поддерживают терминологические базы данных для использования в стандартизации, расположенные по следующим адресам:

- платформа ИСО для онлайн—просмотра материалов по стандартам (Online Browsing Platform, OBP) доступна по адресу https://www.iso.org/obp/ui;
- Электропедия МЭК (IEC Electropedia) доступна по адресу http://www.electropedia.org/.

3.2 Термины, определенные в настоящем документе

3.2.1 **глубокое обучение** (нейронной сети) (deep learning): Подход к созданию обширных иерархических представлений посредством обучения нейронных сетей с большим количеством скрытых слоев.

[ISO/IEC 22989:2022, пункт 3.4.4]

3.2.2 **среда разработки программного обеспечения для глубокого обучения** (deep learning software framework): Инструментарий, который использует набор предварительно созданных и оптимизированных компонентов для определения модели, инкапсулирующей алгоритмы искусственного интеллекта, для вызова данных и использования вычислительных ресурсов.

3.2.3 **модель глубокого обучения** (deep learning model): Алгоритм глубокого обучения, используемый для решения конкретной задачи, обычно относится к информации о структуре вычислительного графа и информации о параметрах, используемых для представления алгоритма глубокого обучения.

3.2.4 **сжатие модели** (model compression): Механизм уменьшения размера модели с помощью таких алгоритмов, как отсечение (прунинг), квантификация, регуляризация, дистилляция знаний и условное начисление.

4 Обозначения и сокращения

ASIC — специализированная интегральная схема (application-specific integrated circuit)

CPU — центральный процессор (central processing unit)

FPGA — программируемая логическая интегральная схема (field-programmable gate array)

GPU — графический процессор (graphics processing unit)

LLVM — низкоуровневая виртуальная машина (low level virtual machine)

NNEF — формат для обмена нейронными сетями (neural network exchange format)

ONNX — открытый обмен нейронными сетями (open neural network exchange)

SGD — стохастический градиентный спуск (stochastic gradient descent)

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 Промышленная реализация среды разработки программного обеспечения для глубокого обучения в области искусственного интеллекта

6.1 Архитектура приложения на основе среды разработки программного обеспечения для глубокого обучения

На рис. 6–1 показана архитектура приложений, основанных на среде разработки программного обеспечения для глубокого обучения.

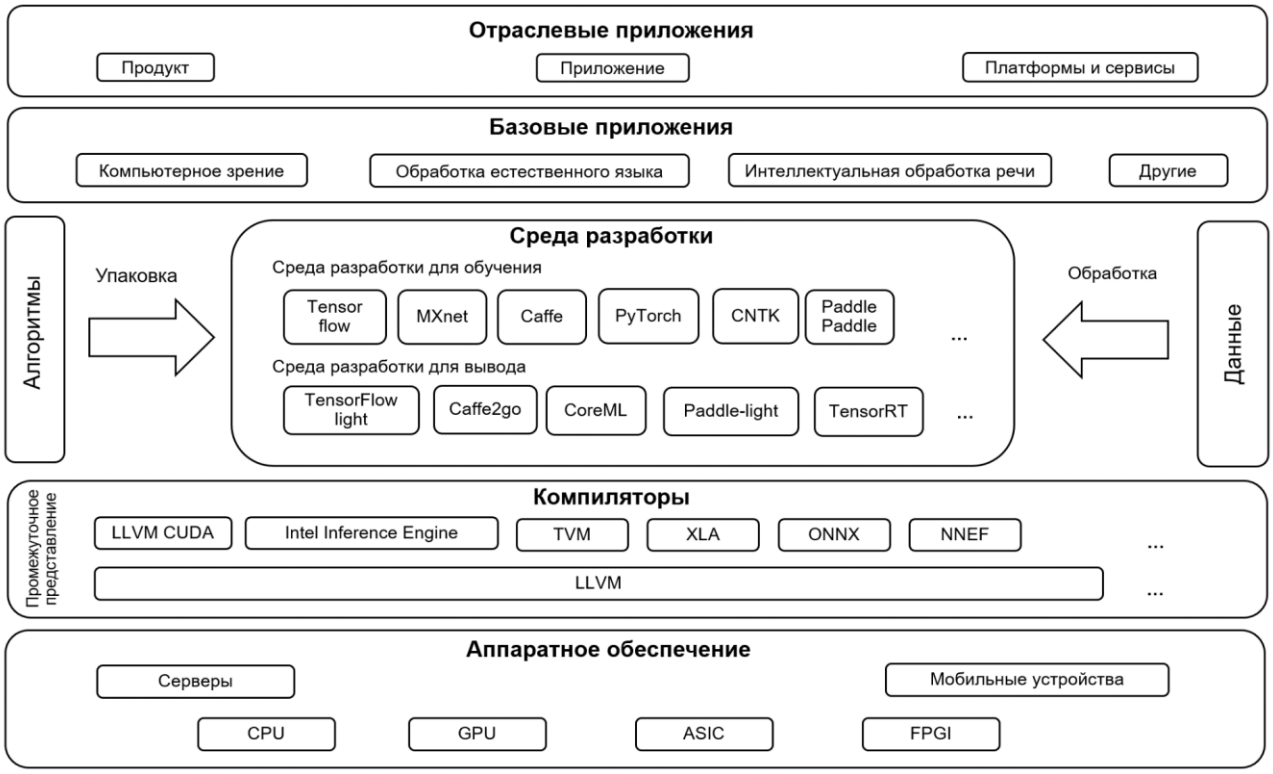

Рисунок 6–1 – Архитектура приложений, основанных на среде разработки программного обеспечения для глубокого обучения

6.2 Аппаратное обеспечение

Под аппаратным обеспечением понимаются центральный процессор (CPU), графический процессор (GPU), специализированная интегральная схема (ASIC), программируемая пользователем матрица логических элементов (FPGA) и другие микросхемы, а также серверы, мобильные устройства и встроенные терминальные устройства, созданные с использованием этих микросхем.

6.3 Компиляторы

Существующие промежуточные представления, включая низкоуровневую виртуальную машину (LLVM) CUDA, Intel Inference Engine, TVM, XLA, открытый обмен нейронными сетями (ONNX), формат обмена нейронными сетями (NNEF) и т. д.

6.4 Среды разработки

Среды разработки используются при инкапсуляции основного аппаратного обеспечения и реализации алгоритмов ИИ, включая среду разработки для обучения и среду разработки для вывода. Среда разработки для обучения включает TensorFlow, PyTorch, Caffe, PaddlePaddle, MXNet и т. д. Среда разработки для вывода включает Tensor RT, TensorFlow Lite, Caffe2go, Paddle Mobile и т. д.

6.5 Базовые приложения

Базовые приложения, как правило, обозначают сценарии применения технологий глубокого обучения в предметных областях искусственного интеллекта таких как компьютерное зрение, обработка естественного языка, интеллектуальная обработка речи.

6.6 Отраслевые приложения

Отраслевые приложения означают применение технологий глубокого обучения к отдельным вертикальным сферам деятельности.

7 Определение требований и методов оценки среды разработки программного обеспечения для глубокого обучения

7.1 Обзор требований

В настоящей Рекомендации отдельно сформулированы соответствующие требования к среде обучения и среде вывода при глубоком обучении. Для среды обучения они охватывают пять аспектов, включая экосистему, простоту использования, производительность, поддерживаемую архитектуру, безопасность и стабильность. Для среды вывода они охватывают четыре аспекта, включая простоту использования, производительность, оптимизацию на уровне аппаратного обеспечения, безопасность и стабильность.

Система показателей для среды обучения включает элементы, перечисленные в Таблице 7–1.

Таблица 7–1 – Система показателей для среды обучения

Система показателей	Показатели
Экосистема	Интерфейс
	Основные разработчики и соисполнители
	Условия, при которых решаются задачи
Простота использования	Построение и конверсия модели
	Вторичная разработка на основе языка высокого уровня
	Пользовательское расширение
	Кроссплатформенность
	Поддержка библиотеки моделей
	Руководства и обучающие материалы
	Динамический и статический графы
	Стабильность
	Возможность отладки
Производительность	Результативность функционирования библиотеки моделей
	Результативность функционирования адаптированной модели
	Поддержка аппаратного ускорения
	Безопасность данных

Окончание Таблицы 7–1	
Поддерживаемая архитектура	CPU/FPGA
	Один GPU/много GPU
	Распределенное обучение
	Поддержка виртуальной среды
Безопасность и стабильность	Использование сторонних библиотек
	Безопасность данных

Система показателей для среды вывода включает элементы, перечисленные в таблице 7–2.

Таблица 7–2 – Система показателей для среды вывода

Система показателей	Показатели
Простота использования	Функциональность оптимизации модели
	Универсальное представление модели
	Кроссплатформенность
Производительность	Скорость вывода
	Время выхода на рабочий режим
	Использование системных ресурсов
	Потребление энергии
Оптимизация на уровне аппаратного обеспечения	Поддержка различного базового аппаратного обеспечения
	Оптимизация набора инструкций
Безопасность и стабильность	Шифрование модели

7.2 Классические модели глубокого обучения

Для оценки среды разработки программного обеспечения [для] глубокого обучения необходимы классические тестовые примеры. В таблице 7–3 перечислены некоторые классические модели глубокого обучения используемые в различных

сценариях тестирования и соответствующие метрики, рекомендуемые при проведении оценки.

Таблица 7–3 – Различные сценарии тестирования и классические модели с соответствующими наборами данных и целями в области качества

Сценарии тестирования	Набор данных	Цели в области качества	Эталонная модель
Классификация изображений	ImageNet	74.90% classification	ResNet-50 v1.5
Обнаружение (облегченных) объектов	COCO 2017	21.2% mAP	SSD (ResNet-34 backbone)
Обнаружение (тяжеловесных) объектов	COCO 2017	0.377 Box min AP, 0.339 Mask min AP	Mask R-CNN
Предварительно обученная модель	Wikipedia	Related to specific tasks	BERT
Перевод (повторяющийся)	WMT English-German	21.8 BLEU	Neural machine translation
Перевод (разовый)	WMT English-German	25.0 BLEU	Transformer
Рекомендация	MovieLens-20M	0.635 HR@10	Neural collaborative filtering
Обучение с подкреплением	Pro games	40.00% move prediction	Mini Go

7.3 Конкретные показатели и методы оценки среды разработки для обучения

7.3.1 Экосистема

7.3.1.1 Интерфейс

Относится к языкам программирования, которые поддерживает среда разработки для обучения.

7.3.1.2 Основные разработчики и соисполнители

Относится к количеству основных разработчиков и соисполнителей, а также активности использования среды разработки для обучения, особенно числу просмотров, звезд, вилок (forks), запросов на вытягивания (pull request), основных разработчиков и соисполнителей в GitHub.

7.3.1.3 Условия, при которых решаются задачи

Относится к механизму и условиях, при которых решаются задачи.

7.3.2 Простота использования

7.3.2.1 Построение и конверсия модели

Относится к согласованности интерфейса прикладного программирования (application programming interface, API), структуры определения модели, формата хранения модели и преобразователей модели.

7.3.2.2 Вторичная разработка на основе языка высокого уровня

Относится к языкам высокого уровня, которые поддерживает среда разработки для обучения. Взвешивает популярность каждого языка для расчета средних значений уровня поддержки языков высокого уровня средой разработки для обучения.

7.3.2.3 Пользовательское расширение

Относится к поддержке и простым в использовании инструментам для разработки расширенных функций, таких как настраиваемые операции, настраиваемый сетевой уровень и добавление строк кода для реализации одной расширенной функциональности.

7.3.2.4 Кроссплатформенность

Относится к совместимости среды разработки для обучения с различными платформами, такими как Linux, Windows, Android, iOS, облачная платформа и т. д.

7.3.2.5 Поддержка библиотеки моделей

Относится к поддерживаемым моделям глубокого обучения, включая CNN, RNN и модели, которые применяются в отраслевых вариантах использования. Необходимо указать количество моделей и охваченные варианты использования.

7.3.2.6 Руководства и обучающие материалы

Должны быть продекларированы руководства и обучающие материалы, в том числе должно быть сказано, существуют ли официальные документы, документы сообщества и материалы от партнерских образовательных учреждений. Если они существуют, необходимо включить количество документов, их полноту, обновления и языки, которые использованы при написании.

7.3.2.7 Динамический граф и статический граф

Относится к механизму выполнения кода среды разработки для обучения, такому как динамический или статический граф.

7.3.2.8 Стабильность

Относится к механизму повышения стабильности среды для обучения.

7.3.2.9 Возможность отладки

Относится к гибкости и эффективности механизмов отладки.

7.3.3 Производительность

7.3.3.1 Эксплуатационная производительность библиотеки моделей

Относится ко времени исполнения и к использованию памяти при выполнении классических моделей из официальной библиотеки моделей с различными размерами пакетов.

7.3.3.2 Эксплуатационная производительность адаптированной модели

Относится к параллелизму, стабильности и количеству независимых параметров. Для параллелизма рекомендуется проводить оценки, чтобы получить

конкретные расчетные величины, выполняемые одновременно, и интерактивную эффективность среды разработки. Для стабильности рекомендуется проводить оценки для получения статистических данных о сбоях в течение недели путем непрерывного выполнения задач на основе среды для обучения. Для количества независимых параметров рекомендуется проводить оценки, чтобы получить наибольшую пропускную способность среды для обучения и выровнять ее с его возможностями.

7.3.3.3 Поддержка аппаратного ускорения

Относится к оптимизации среды разработки для обучения при использовании различного базового аппаратного обеспечения, такого как CPU, GPU, FPGA, специализированные ускорители ИИ и т. д.

7.3.4 Поддерживаемая архитектура

7.3.4.1 CPU/FPGA

Относится к производительности классических моделей, исполняемых на CPU или FPGA, включая время обучения и использование памяти.

7.3.4.2 Один GPU/много GPU

Относится к производительности синхронизированного стохастического градиентного спуска (SGD) классических моделей и других сложных методов оптимизации моделей, выполняемых на одном GPU и нескольких GPU.

7.3.4.3 Распределенное обучение

Относится к синхронизированному поведению SGD классических моделей и другим сложным методам оптимизации моделей, выполняемым на одной и той же машине и на разных машинах.

7.3.4.4 Поддержка виртуальной среды

Относится к применимости среды разработки для обучения к использованию в виртуальной среде, на виртуальной машине и т. д.

7.3.4.5 Поддержка операционных систем

Относится к использованию среды разработки для обучения в нескольких операционных системах, таких как Linux, Windows и macOS.

7.3.5 Безопасность и стабильность

7.3.5.1 Использование сторонней библиотеки

Относится к сторонней библиотеке, используемой в среде разработки для обучения.

7.3.5.2 Безопасность данных

Относится к тому, поддерживаются ли средой разработки для обучения зашифрованные данные при обучении и предоставляется ли инструмент шифрования.

7.4 Конкретные показатели и методы оценки среды разработки для вывода

7.4.1 Простота использования

7.4.1.1 Функциональность оптимизации модели

Относится к возможности сжатия модели, ускорения компонентов алгоритма и компонентов оптимизации гиперпараметров.

7.4.1.2 Представление универсальной модели

Относится к тому, поддерживает ли среда разработки для вывода универсальный формат представления модели глубокого обучения, например ONNX.

7.4.1.3 Кроссплатформенность

Относится к совместимости среды разработки для вывода с различными платформами, такими как Linux, Windows, Android, iOS, облачная платформа и т. д.

7.4.2 Производительность

7.4.2.1 Скорость вывода

Относится ко времени вывода при исполнении классических моделей глубокого обучения с использованием среды разработки для вывода при различном потреблении энергии.

7.4.2.2 Скорость выхода на рабочий режим

Относится ко времени выхода на рабочий режим при исполнении классических моделей глубокого обучения с использованием среды разработки для вывода на различных массивах данных.

7.4.2.3 Использование системных ресурсов

Относится к потреблению памяти, использованию памяти центрального процессора, использованию памяти графического процессора при исполнении моделей глубокого обучения разных масштабов.

7.4.2.4 Потребление энергии

Относится к затратам энергии за один час за счет выполнения моделей глубокого обучения разных масштабов.

7.4.3 Оптимизация на уровне аппаратного обеспечения

7.4.3.1 Поддержка различного аппаратного обеспечения

Относится к способности среды разработки для вывода поддерживать различное аппаратное обеспечение, такое как FPGA, ASIC, GPU и т. д.

7.4.3.2 Оптимизация набора инструкций

Относится к оптимизации поддерживаемого средой разработки для вывода набора инструкций, такого как усовершенствованная RISC-машина (ARM), набор инструкций INTEL.

7.4.4 Безопасность и стабильность

7.4.4.1 Шифрование модели

Относится к стратегии шифрования обученной модели вывода в среде для разработки вывода.

Приложение ДА
(справочное)
Сведения о соответствии ссылочных национальных стандартов ссылочным международным стандартам, использованным в качестве ссылочных в примененном международном документе

Таблица ДА.1

Обозначение ссылочного международного стандарта	Степень соответствия	Обозначение и наименование соответствующего национального стандарта
ISO/IEC 22989:2022, Information technology — Artificial intelligence — Artificial intelligence concepts and terminology	NEQ*	ПНСТ 553–2021** Информационные технологии. Искусственный интеллект. Концепции искусственного интеллекта и терминология (ISO/IEC DIS 22989, NEQ)
* В настоящей таблице использовано следующее условное обозначение степени соответствия стандарта: NEQ — неэквивалентный стандарт. ** В России принят предварительный национальный стандарт, основанный на ранней редакции международного стандарта		

ПНСТ
(проект)

УДК 004.01:006.354 ОКС 35.020

Ключевые слова: среда разработки программного обеспечения для глубокого обучения, простота использования, экосистема, оценка, производительность, безопасность и стабильность, поддерживаемая архитектура, оптимизация на уровне аппаратного обеспечения

Руководитель разработки

Председатель совета директоров
Института развития информационного общества,
академик Российской инженерной академии Ю. Е. Хохлов