**Программное обеспечение**

**«Neural Processing Service»**

**(ПО «NPS»)**

**Описание системы**

Листов

Москва 2022

1. ОГЛАВЛЕНИЕ

[ОГЛАВЛЕНИЕ 2](#_Toc115366669)

[ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИй 3](#_Toc115366670)

[ВВЕДЕНИЕ 4](#_Toc115366671)

[1. Общие сведения и логика работы системы 5](#_Toc115366672)

[2. Функциональные характеристики 7](#_Toc115366673)

[2.1 Функции алгоритмического сервиса 7](#_Toc115366674)

[2.2 Функции промежуточного сервиса 7](#_Toc115366675)

[2.3 Функции базы данных 7](#_Toc115366676)

[2.4 Функции сервиса аналитики 7](#_Toc115366677)

[3. Программное и техническое обеспечение 8](#_Toc115366678)

1. ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИй

| Сокращение | Детальная расшифровка |
| --- | --- |
| БД | База данных |
| ПО | Программное обеспечение |
| СУБД | Система управления базами данных |
| GPU | Graphics Processing Unit |

# ВВЕДЕНИЕ

В данном документе приведены сведения о функциональных характеристиках, логической структуре и функционировании программного обеспечения «Neural Processing Service» (далее Система), а также требования к техническому обеспечению.

Обслуживающий персонал должен иметь высшее техническое образование и квалификацию уровня инженер и выше.

1. Общие сведения и логика работы системы

Программное обеспечение предназначено для пакетной распределенной обработки входных данных нейросетевыми алгоритмами с оптимизацией под конкретные аппаратные платформы.

Программное обеспечение NPS состоит из пяти основных сущностей:

* Веб-сервис, обеспечивающий доступ пользователям системы.
* Промежуточный сервис, отвечающий за получение запросов от клиентских приложений и перенаправление их алгоритмическому сервису, а также за хранение данных в системе управления базами данных (СУБД).
* СУБД, в которой хранятся данные.
* Алгоритмический сервис, позволяющий запускать различные алгоритмы диагностики на основе как нейронных сетей глубокого обучения, так и классических алгоритмов машинного обучения.
* Сервис аналитики, отвечает за препроцессинг и постпроцессинг данных, извлечение метаданных из изображений, привязку изображений к различным категориям, формирование метаданных после работы алгоритмов машинного обучения.

Внешнее приложение или сервис запускает процесс обработки специальной командой. При этом статус обработки меняется на «analysis», сервис аналитики начинает препроцессинг. На этом этапе производится извлечение метаданных из входящих изображений и привязка данных к категориям (настраиваются отдельно при настройке системы). После препроцессинга статус меняется на «analyzing», при этом промежуточный сервис создаёт задание, состоящее из подзаданий, соответствующих одному конкретному алгоритму машинного обучения. В процессе обработки каждого подзадания запускается алгоритмический сервис, получающий пакеты соответствующих изображений и настройки соответствующего алгоритма. Полученные в результате работы алгоритмов машинного обучения метаданные записываются в БД системы.

После обработки всех подзаданий, процесс обработки получает статус постпроцессинга. При этом запускается сервис аналитики. На этом этапе происходит оценка полученных на предыдущих этапах метаданных и формирование дефектов на основе этой оценки и уставок, имеющихся в БД системы. Статус обработки меняется на «completed» и результаты высылаются на клиентское приложение и сохраняются в БД.

Стоит отметить, что сервис в рамках задания позволяет запускать на одних данных несколько последовательных алгоритмов, комбинируя их в более комплексный алгоритм анализа данных.

1. Функциональные характеристики
	1. Функции алгоритмического сервиса

Алгоритмический сервис отвечает за обработку входящих данных путём запуска соответствующих алгоритмов машинного обучения на основе заданий и настроек алгоритмов, формируемых промежуточным сервисом.

* 1. Функции промежуточного сервиса

Промежуточный сервис отвечает за управление остальными сервисами системы, формирование задач, подзадач, настроек алгоритмов машинного обучения, передачу данных между сервисами системы.

* 1. Функции базы данных

База данных служит для хранения данных системы, данных синхронизации сервисов системы, хранение настроек алгоритмов, других служебных данных системы.

* 1. Функции сервиса аналитики

Сервис аналитики, отвечает за препроцессинг и постпроцессинг данных, извлечение метаданных из изображений, привязку изображений к классам, формирование выходных метаданных после работы алгоритмов машинного обучения.

1. Программное и техническое обеспечение

В силу использования в задачах обнаружения дефектов алгоритмов, основанных на глубоких нейронных сетях, сервис, ответственный за запуск данных алгоритмов, достаточно требователен к вычислительным ресурсам.

Новые видеокарты компании Nvidia на архитектуре «Turing» не просто привносят очередное повышение производительности, но и содержат в себе ряд технологических инноваций и являются первыми игровыми решениями, которые поддерживают трассировку лучей в реальном времени. Поэтому даже название GeForce GTX было изменено на GeForce RTX. В семействе «Turing» можно выделить несколько ключевых изменений: абсолютно новая архитектура GPU, появление новых вычислительных блоков – тензорных и RT (ray-tracing) ядер, ускоренная обработка шейдеров.

При тестировании в системах искусственного интеллекта оказалось, что наличие тензорных ядер критически сказывается на производительности решения в целом, ускоряя прямой проход (inference) нейронных сетей на 20-30%. Поэтому в качестве графических ускорителей в аппаратной платформе для алгоритмического сервиса рекомендуется использовать видеокарты Nvidia поколения RTX.

Минимальные аппаратные требования для центрального сервера испытательного стенда с установленными сервисами системы следующие:

• центральный процессор Intel Xeon Silver 1 поколения или выше с не менее чем 8 аппаратными или 16 виртуальными ядрами;

• видеокарта Nvidia Quadro RTX 4000 или лучше для отработки нейросетевых алгоритмов;

• не менее 64Гб оперативной памяти;

• файловое хранилище для СУБД в системах RAID 5, 6 или 10;

• не менее 400Гб свободного места на диске;

• сетевое соединение не менее 1Гбит/сек.

Однако в случае тестирования испытательного стенда программное обеспечение может быть установлено и на более слабый вычислительный блок, включая игровой персональный компьютер:

• центральный процессор Intel Core i7 6 поколения или выше с не менее чем 4 аппаратными или 8 виртуальными ядрами (данное требование является минимальным для запуска фреймворка Intel OpenVino, начиная с версии 2021.4 LTS);

• видеокарта GeForce GTX 1070 или лучше для отработки нейросетевых алгоритмов (данное требование по поколению является минимальным для запуска фреймворка Nvidia TensorRT, начиная с версии 7.x, также для некоторых нейросетевых алгоритмов требуется существенный объем памяти, поэтому минимальные рекомендации – от 6Гб и выше, у карты GTX 1070 8Гб оперативной памяти на борту);

• не менее 16Гб оперативной памяти.

Минимальные программные требования для центрального сервера с установленными сервисами системы следующие:

• ОС семейства GNU/Linux (Ubuntu Linux 20.04 и новее, РЕД ОС и т.п.);

• Установленные драйверы GPU nVidia последней версии

• система управления базами данных PostgreSQL;

• выход в интернет с выделенным IP-адресом.