

КОНЦЕПЦІЯ ІНТЕГРАЦІЇ ПОТОКІВ ВЕЛИКИХ ДАНИХ З МОДЕЛЯМИ ГЛИБОКОГО НАВЧАННЯ

Ілля Черняк

CONCEPT OF INTEGRATION OF BIG DATA STREAMS WITH DEEP LEARNING MODELS

Інтеграція потоків великих даних з моделями глибокого навчання полягає у створенні системи, яка здатна об'єднати дані з різних джерел, включаючи як пакетні дані, так і дані в режимі реального часу. Основні кроки цього процесу включають збір та попередню обробку даних, паралельне навчання глибоких нейронних мереж, а також можливість використовувати моделі для аналізу нових даних в режимі реального часу [1]. Концептуальні основи для такої системи включають наступні ключові аспекти:

1. Збір та вибір даних: система повинна забезпечити можливість вибору даних з різних джерел, включаючи пакетні дані та дані в режимі реального часу. Це може включати в себе інтеграцію з різними джерелами даних, такими як Apache Hadoop, Apache Spark, Apache Storm, бази даних, API, тощо.

2. Попередня обробка даних: система повинна надавати інструменти для попередньої обробки даних, включаючи очищення, нормалізацію, видалення аномалій та відновлення відсутніх даних. Ця операція дозволяє підготувати дані для подальшого навчання моделей глибокого навчання.

3. Навчання глибоких нейронних мереж: система повинна підтримувати навчання моделей глибокого навчання на оброблених даних. Це може включати в себе розпаралелення на рівні даних та моделі, використання еволюційного підходу або інших методів навчання глибоких нейронних мереж.

4. Управління експериментами: система повинна дозволяти користувачам проводити декілька експериментів одночасно та вести їх моніторинг. Це допомагає в розвідуванні різних підходів та порівнюванні результатів.

5. Обробка різних типів даних: система повинна бути гнучкою і підтримувати обробку різних типів даних, включаючи текстові, зображення, аудіо, числові дані та інші.

6. Автоматичне виконання та відтворення кроків: система повинна мати можливість виконувати операції у фоновому режимі та записувати виконані кроки для подальшого відтворення на великому обсязі даних.

7. Оцінка якості даних і відновлення відсутніх даних: система повинна надавати інструменти для оцінки якості вхідних даних та відновлення відсутніх даних, щоб забезпечити якісну підготовку даних для моделей глибокого навчання.

8. Робота в режимі реального часу та на обмежених ресурсах: система повинна бути здатною працювати в режимі реального часу для поточкових даних та на системах з обмеженими обчислювальними ресурсами.

9. Моделювання нейронних мереж: система повинна дозволити користувачам моделювати різні архітектури нейронних мереж в залежності від поставлених завдань та вимог до ефективності та достовірності.

Література

1. Комар М.П. Інформаційна технологія інтелектуальної обробки та аналізу великих даних. Вісник Хмельницького національного університету. Технічні науки. 2020. № 5. С. 125–130.