УДК 004.89, 004.6

І.А. Ляпандра, В.В. Івахів, В.С. Білоус
Західноукраїнський національний університет, Україна

МЕТОДИ ТА ЗАСОБИ ОБРОБКИ ВЕЛИКИХ ДАНИХ

I.A. Liapandra, V.V. Ivakhiv, V.S. Bilous

METHODS AND TOOLS FOR BIG DATA PROCESSING

Експоненційний ріст цифрових даних, що генеруються з численних і різних джерел, робить неможливим зберігання, обробку та аналіз за допомогою традиційних методів. Ці обмеження призвели до еволюції технологій великих даних. Великі дані, які визначаються швидким зростанням обсягу, різноманітності та швидкості даних, як правило, мають справу з неструктурованими даними, які потребують великого пакетного аналізу або аналізу в реальному часі. Щоб отримати значущі результати з таких обсягів даних, необхідна величезна потужність з точки зору сховищ і обчислювальних ресурсів, а також потрібні методи паралельної обробки даних.

Глибоке навчання, підмножина машинного навчання, – це техніка, яка використовується для аналізу та обробки величезної кількості даних, щоб знайти абстрактні й корисні моделі. Якщо застосувати глибоке навчання до великих даних, то можна знайти невідомі й корисні закономірності, які неможливо знайти на основі традиційних методів [1, 2]. Традиційні підходи машинного навчання показують кращу продуктивність для меншої кількості вхідних даних. Оскільки обсяг даних перевищує певну кількість, продуктивність традиційних підходів машинного навчання стає стабільною, тобто досягає плато. Однак продуктивність підходів глибокого навчання зростає по відношенню до збільшення кількості даних. Тому використання глибокого навчання є перспективним щодо великих даних [3-5].

Великі дані потребують нових і складних алгоритмів, заснованих на технологіях машинного навчання та глибокого навчання, щоб обробляти дані в режимі реального часу з високою точністю та ефективністю [6]. Методи глибокого навчання надали потужні інструменти для роботи з великими обсягами даних, оскільки вони витягують з них функції вищого рівня для отримання ієрархічних представлень.

Глибоке навчання отримало застосування у різних областях, зокрема розпізнавання мовлення, акустичного моделювання для класифікації звуку, обробки зображень, таких як рукописна класифікація, класифікація сцен дистанційного зондування високої роздільної здатності, обробка природної мови, комп’ютерний зір, розпізнавання образів тощо [7].

Зі збільшенням великої кількості даних, що генеруються з різних джерел, існуючі технології обробки даних не придатні для цього. Фреймворки великих даних допомагають зберігати, аналізувати та обробляти такі величезні обсяги даних [6].

Аналіз методів та засобів обробки великих даних показав відсутність програмного середовища для інтеграції систем обробки великих даних та моделей глибокого навчання. Використовуючи таке середовище, користувачі зможуть трансформувати дані, що надходять з різних потоків великих даних, у формат, необхідний для навчання моделей глибокого навчання. Таке середовище забезпечить структуру, яка інтегрує різні потоки великих даних та моделі глибокого навчання, а також дозволить користувачам виконувати різні маніпуляції з потоками даних. Користувачі зможуть швидко побудувати та виконати різні експерименти з підмножиною набору даних, щоб отримати з нього значущі результати.

Інтегроване середовище повинно надавати користувачеві можливість вибору даних (пакетних даних / даних у реальному часі) з різних архітектур великих даних, попередньої обробки даних для перетворення даних у необхідний формат, а потім навчання моделей глибокого навчання з використанням попередньо оброблених даних .

Інтегроване середовище повинно складатися з наступних компонентів:

1) інтерфейс користувача: надаватиме користувальницький інтерфейс, який буде дозволяти користувачам вибирати та виконувати різні операції.

2) контролер: оброблятиме різні компоненти архітектури.

3) блок вибору даних: дозволятиме вибирати різні великі потоки даних.

4) блок попередньої обробки даних: виконуватиме різні операції попередньої обробки даних.

5) Блок навчання даних: виконуватиме навчання моделей глибокого навчання.

Інтегроване середовище повинно забезпечувати виконання наступних функцій:

1) підтримка пакетних даних від Apache Hadoop та потоків даних у режимі реального часу від Apache Spark та Apache Storm;

2) підтримка навчання моделей глибоких нейронних мереж на основі еволюційного підходу на системах з одним або декількома графічними процесорами на рівні розпаралелення даних та на рівні розпаралелення моделі;

3) підтримка одночасного проведення декількох експериментів одночасно, щоб забезпечити більш швидке дослідження проблемної області;

4) підтримка обробки декількох типів даних, присутніх у наборі даних;

5) надання можливості автоматично виконувати всі кроки користувача у фоновому режимі на більшому наборі даних без будь-якого втручання користувача;

6) надання можливості записування всіх виконаних кроків, а згодом дозволяє користувачеві запускати їх на всьому наборі даних;

7) надання можливості виконання операцій попередньої обробки даних, оцінити якість вхідних даних.

Література:

1. M. Gheisari, G. Wang, M. Z. A. Bhuiyan. A survey on deep learning in big data. 2017. Vol. 2. Pp. 173–180.

2. Wang C., Shakhovska N., Sachenko A., Komar M. A New Approach for Missing Data Imputation in Big Data Interface. Information Technology and Control. 2020. Vol. 49. No 4. Pp. 541-555.

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

4. Комар М.П. Хорунжий О.В., Лічак В.М., Бучинський Р.З. Аналіз та обробка великих даних на основі глибоких нейронних мереж. Актуальні задачі сучасних технологій : зб. тез доп. міжнар. наук.-техн. конф., Тернопіль, 28-29 листопада, 2018. Т.2. С. 86.

5. Комар М.П. Перевізник Р.М., Неспляк Д.Б. та ін. Проектування прикладних систем обробки та аналізу великих даних на основі глибоких нейронних мереж. Актуальні задачі сучасних технологій : зб. тез доп. міжнар. наук.-техн. конф., Тернопіль, 25-26 листопада, 2020. Т.2. С. 30-31.

6. W. Inoubli, S. Aridhi, H. Mezni, M. Maddouri, E.M. Nguifo. An Experimental Survey on Big Data Frameworks. Future Generation Computer Systems. [Електронний ресурс]. – Режим доступу: https://arxiv.org/pdf/1610.09962.pdf.

7. Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. Neurocomputing. 2017. Vol. 234. Pp.11-26.