УДК 004.048

Бобко О.Л., студент 2го курсу магістерського рівня спеціальності 121 «Інженерія програмного забезпечення» ВНТУ «ФІТКІ»
Рейда О.М. к.т.н., доцент кафедри ПЗ

ВИКОРИСТАННЯ CPU, GPU, TPU ДЛЯ НАВЧАННЯ НЕЙРОННИХ МЕРЕЖ

Вінницький національний технічний університет

Анотація

У даній роботі було досліджено різні типи обчислювальних систем які використовуються при навчанні нейронних мереж. Процес тренування моделі передбачає надзвичайно велику кількість математичних операцій, тому час навчання потенційно може тривати дуже довго.

Ключові слова: CPU, GPU, TPU, FNN, CNN, RNN

Abstract

In this work, various types of computing systems used in training neural networks were investigated. The process of training a model involves an enormous number of mathematical operations, hence the training time potentially can take a very long time.

Keywords: CPU, GPU, TPU, FNN, CNN, RNN

Вступ

Навчання моделі нейронної мережі це процес який вимагає дуже великої кількості обчислювальних операцій. Саме тому необхідно розуміти потенційне обчислювальне навантаження при навчанні, та час навчання. Визначення необхідного обладнання також прямо впливає на фінансові затрати на тренування моделі [1].

Результати дослідження

Підчас машинного навчання виконується певний алгоритм, що отримує навчальні дані на базі яких будує модель яка здатна виконувати певні дії, наприклад прогнозування. Зрештою модель набуває здатності робити припущення на основі реальних даних, наприклад визначати та класифікувати об’єкт на зображенні.

Процес навчання потребує використання таких ресурсів як: процесорний час, об’єм тренувальних даних, загальний час тренування. Визначення або передбачення необхідної кількості ресурсів залежить від таких факторів як залежність витраченого процесорного часу до об’єму тренувальних даних та загальному часу навчання, так і від таких факторів як наприклад типу навчання (з учителем, без учителя), архітектура нейронної мережі (FNN, CNN, RNN), тип обчислювальної системи, тобто CPU, GPU, або TPU [2].

Відповідно до теми дослідження розглянемо який варіант обчислювальної системи підходить найкраще і коли. Розглянемо їх загальні властивості:

CPU (Central Processing Unit) – це основний обчислювальний пристрій комп'ютера, відповідальний за виконання програм, обробку даних та керування іншими компонентами. Він підтримує широкий спектр завдань і використовується для загального призначення: запуску операційних систем, веб-браузерів, офісних програм тощо. CPU зазвичай має декілька ядер, кожен з яких може виконувати окремі завдання.

GPU (Graphics Processing Unit) – спеціалізується на обробці графіки та обчисленнях паралельних завдань. Використовується в графіці, відеоіграх, наукових обчисленнях, штучному

інтелекті, машинному навчанні та обробці великих обсягів даних. GPU має значно більше ядер, ніж CPU, і дозволяє швидше виконання паралельних завдань.
TPU (Tensor Processing Unit) – спеціалізований обчислювальний пристрій, розроблений Google для прискорення обчислень у машинному навчанні та штучному інтелекті. Він оптимізований для роботи з тензорами та матрицями, які використовуються в нейронних мережах. TPU зазвичай має велику швидкість та ефективність у порівнянні з CPU та навіть GPU у завданнях машинного навчання [3].

Отже, CPU, як пристрій загального призначення не має достатньої кількості ядер для ефективної. Наприклад на час виконання даного дослідження найбільшу кількість ядер (128) мають процесори сімейства AMD EPYC Milan 7003 Series. У той час коли NVIDIA GEFORCE RTX 4060 GPU – підтримує до 3072 ядер, що оптимізовані під виконання математичних операцій. Тому очевидно, що для навчання варто використовувати GPU або TPU.

При виборі між GPU та TPU остаточне рішення зазвичай залежить від конкретних вимог проекту – обмежень бюджету, можливостей середовища розробки і т.д. Наприклад, якщо точність має значення, то може бути відданий перевагу GPU, оскільки вони пропонують більшу гнучкість у використанні точності, у той час як у разі, якщо час від моменту створення моделі до її впровадження є важливим, тоді TPU може бути відданий перевагу, оскільки вони забезпечують швидший час інференсу порівняно з їхніми аналогами – GPU, при цьому забезпечуючи покращені переваги енергоефективності на масштабах протягом тривалого часу.

Висновки

У результаті проведеного дослідження було визначено що, кожен тип пристрою підходить для навчання нейронних мереж різної складності. Однак, використання CPU для тренування моделей вважається не ефективним на великих наборах навчальних даних. Враховуючи потенційно великі затрати на закупівлю обладнання (GPU) вважається більш ефективним використовувати хмарні технології, що надають можливість використовувати різні типи процесорів на вибір в залежності від обчислювального навантаження. Узагальнюючи можна сказати, що CPU використовується для загального призначення, GPU – для паралельних обчислень та графіки, а TPU – для швидкого та ефективного виконання завдань у сфері машинного навчання та обробки даних.

Перелік джерел посилання

1. Yuqi Li. How to Estimate the Time and Cost to Train a Machine Learning Model. URL: https://towardsdatascience.com/how-to-estimate-the-time-and-cost-to-train-a-machine-learning-model-eb6c8d433ff7 (дата звернення: 30.11.2023).
2. Edward Hu, Greg Yang, Jianfeng Gao. µTransfer: A technique for hyperparameter tuning of enormous neural networks. URL: https://www.microsoft.com/en-us/research/blog/%C2%B5transfer-a-technique-for-hyperparameter-tuning-of-enormous-neural-networks/ (дата звернення: 30.11.2023)
3. Brian Mathew. TPU vs GPU vs CPU: Understanding the difference. URL: https://computertechnicians.com.au/tpu-vs-gpu-vs-cpu/ (дата звернення: 30.11.2023).