Demystifying Parallel and Distributed Deep Learning: An In-Depth Concurrency Analysis:

Làm sáng tỏ Học tập sâu song song và phân tán: Phân tích đồng thời trong chiều sâu

Chart, bar chart

Description automatically generated

Fig. 4a shows a breakdown of the number of nodes used in deep learning research over the years. It started very high with the large-scale DistBelief run, reduced slightly with the introduction of powerful accelerators and is on a quick rise again since 2015 with the advent of large-scale deep learning. Out of the 240 reviewed papers, 73 make use of distributed-memory systems and provide details about their hardware setup. We observe that large-scale setups, similar to HPC machines, are commonplace and essential in today’s training

Hình 4a cho thấy sự phân tích về số lượng các nút được sử dụng trong nghiên cứu học tập sâu trong những năm qua. Nó bắt đầu rất cao với việc chạy DistBelief quy mô lớn, giảm nhẹ với sự ra đời của các máy gia tốc mạnh mẽ và đang tăng nhanh trở lại kể từ năm 2015 với sự ra đời của học sâu quy mô lớn. Trong số 240 bài báo được đánh giá, 73 bài báo sử dụng hệ thống bộ nhớ phân tán và cung cấp thông tin chi tiết về thiết lập phần cứng của chúng. Chúng tôi nhận thấy rằng các thiết lập quy mô lớn, tương tự như máy HPC, là phổ biến và cần thiết trong đào tạo ngày nay

Fig. 4b shows a breakdown of the different communication mechanisms that were specified in 55 of the 73 papers using multi-node parallelism. It shows how the community quickly recognized that deep learning has very similar characteristics than large-scale HPC applications. Thus, beginning from 2016, the established MPI interface became the de-facto portable communication standard in distributed deep learning.

Hình 4b cho thấy sự phân tích của các cơ chế giao tiếp khác nhau đã được chỉ định trong 55 trong số 73 bài báo sử dụng song song nhiều nút. Nó cho thấy cách cộng đồng nhanh chóng nhận ra rằng học sâu có những đặc điểm rất giống với các ứng dụng HPC quy mô lớn. Do đó, bắt đầu từ năm 2016, giao diện MPI được thiết lập đã trở thành tiêu chuẩn giao tiếp di động thực tế trong học sâu phân tán.

Chart

Description automatically generated

Parallel Algorithms

 For example, assuming we can process one operation per time unit, then the time needed to process the graph on a single processor is T1 = W and the time needed to process the graph on an infinite number of processes is T∞ = D. The average parallelism in the computation is W/D, which is often a good number of processes to execute the graph with. Furthermore, we can show that the execution time of such a DAG on p processors is bounded by

Hai tham số này cho phép chúng ta mô tả độ phức tạp tính toán trên một hệ thống song song. Ví dụ: giả sử chúng ta có thể xử lý một thao tác trên một đơn vị thời gian, thì thời gian cần thiết để xử lý đồ thị trên một bộ xử lý là T1 = W và thời gian cần thiết để xử lý đồ thị trên vô số quá trình là T∞ = D. độ song song trung bình trong tính toán là W / D, thường là một số quá trình tốt để thực hiện đồ thị với. Hơn nữa, chúng tôi có thể chỉ ra rằng thời gian thực thi của một DAG như vậy trên bộ xử lý p bị giới hạn bởi

