

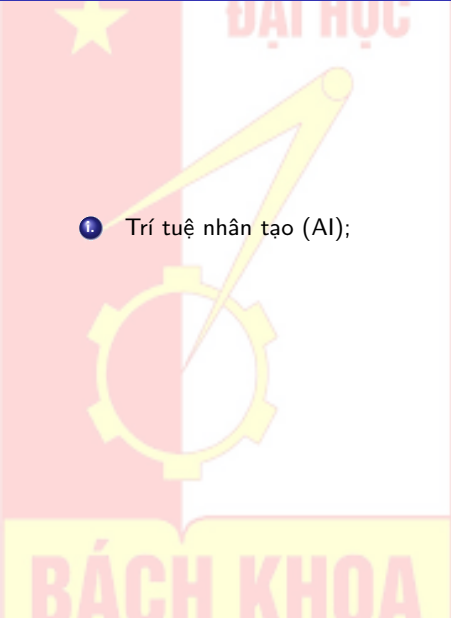
TÍNH TOÁN SONG SONG HỆ THỐNG HỌC SÂU PHÂN TÁN

Sinh viên thực hiện: Nguyễn Hữu Thuật

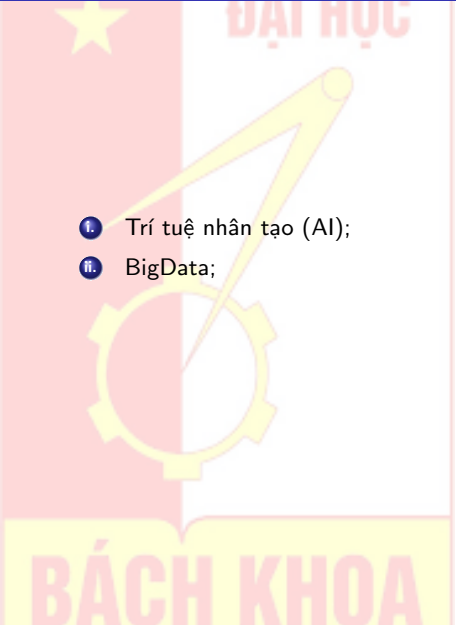
Viện toán ứng dụng và tin học
Đại học bách khoa Hà Nội

Tháng 08 năm 2021

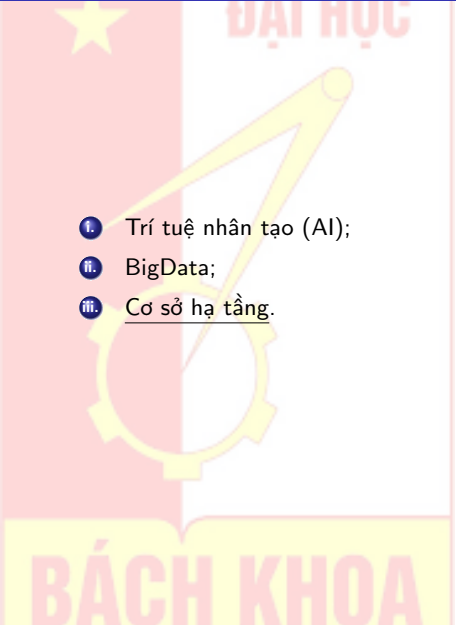
Lý do chọn đề tài

- 
1. Trí tuệ nhân tạo (AI);

Lý do chọn đề tài

- 
- i. Trí tuệ nhân tạo (AI);
 - ii. BigData;

Lý do chọn đề tài

- 
- i. Trí tuệ nhân tạo (AI);
 - ii. BigData;
 - iii. Cơ sở hạ tầng.

Cấu trúc

Cấu trúc chính của bài tiểu luận gồm 03 phần:

- i. Giới thiệu một số lý thuyết cơ sở;
- ii. Giới thiệu về học sâu phân tán và những vấn đề liên quan;
- iii. Áp dụng thực tế và kết quả tích lũy.



ĐẠI HỌC



BÁCH KHOA

Phần 1: Giới thiệu một số lý thuyết cơ sở

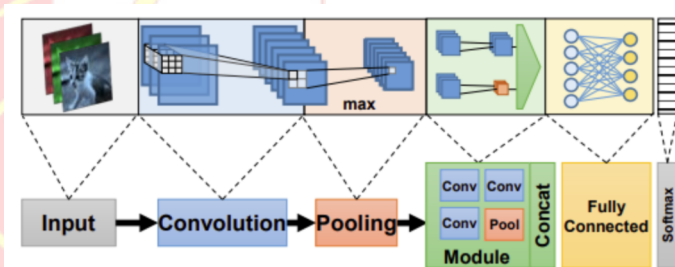
Lý thuyết cơ sở

Học sâu (Deep Learning):

Lý thuyết cơ sở

Học sâu (Deep Learning):

- Mô hình dữ liệu trừu tượng hóa;
- Sử dụng nhiều lớp;
- Biến đổi phi tuyến.



Deep Network

Lý thuyết cơ sở

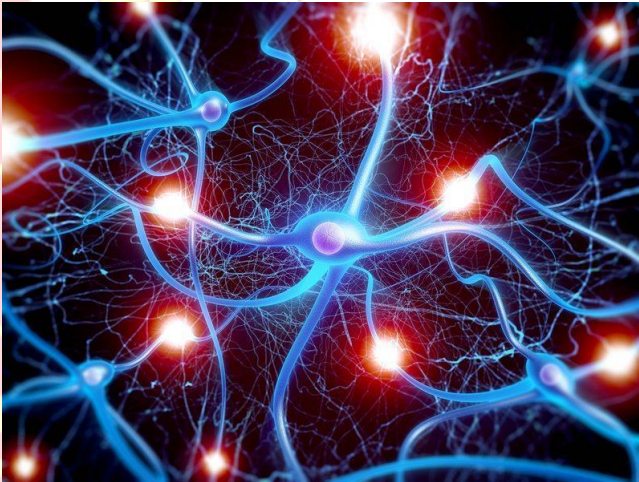
Nơ-ron:



Lý thuyết cơ sở

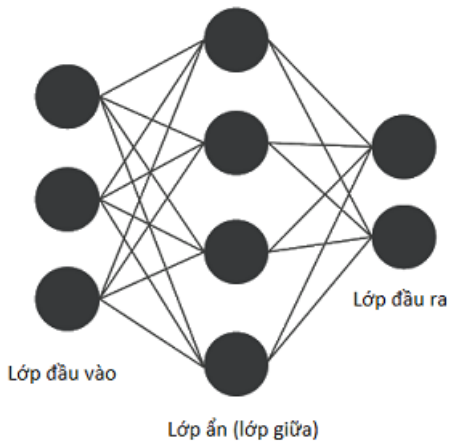
Nơ-ron:

- Mô phỏng theo não người;
- Sử dụng hàm kích hoạt.



Lý thuyết cơ sở

Mạng Nơ-ron:



Các yếu tố trong huấn luyện mô hình Mạng nơ-ron:

Lý thuyết cơ sở

Các yếu tố trong huấn luyện mô hình Mạng nơ-ron:

- Dữ liệu:

Các yếu tố trong huấn luyện mô hình Mạng nơ-ron:

- Dữ liệu:

- Dữ liệu lớn (Bigdata):

- i. Dung lượng;
- ii. Tính đa dạng;
- iii. Vận tốc;
- iv. Tính xác thực.

Lý thuyết cơ sở

Các yếu tố trong huấn luyện mô hình Mạng nơ-ron:

- Dữ liệu:
 - Dữ liệu lớn (Bigdata):
 - i. Dung lượng;
 - ii. Tính đa dạng;
 - iii. Vận tốc;
 - iv. Tính xác thực.
- Mô hình;

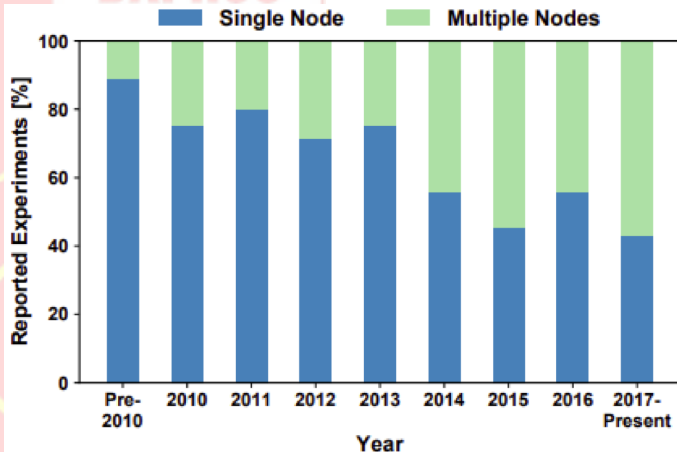
Các yếu tố trong huấn luyện mô hình Mạng nơ-ron:

- Dữ liệu:
 - Dữ liệu lớn (Bigdata):
 - i. Dung lượng;
 - ii. Tính đa dạng;
 - iii. Vận tốc;
 - iv. Tính xác thực.
- Mô hình;
- Hàm mục tiêu;

Các yếu tố trong huấn luyện mô hình Mạng nơ-ron:

- Dữ liệu:
 - Dữ liệu lớn (Bigdata):
 - i. Dung lượng;
 - ii. Tính đa dạng;
 - iii. Vận tốc;
 - iv. Tính xác thực.
- Mô hình;
- Hàm mục tiêu;
- Thuật toán tối ưu.

Kiến trúc máy tính song song



Training with Single vs. Multiple Nodes

Kiến trúc máy tính song song

Máy tính song song đơn (Single-machine Parallelism)

Kiến trúc máy tính song song

Máy tính song song đơn (Single-machine Parallelism)

- Đa tiến trình (multiple processes);
- Đa luồng (multiple threads);
- Kết hợp cả hai.

Kiến trúc máy tính song song

Đa máy tính song song (multi-machine Parallelism))

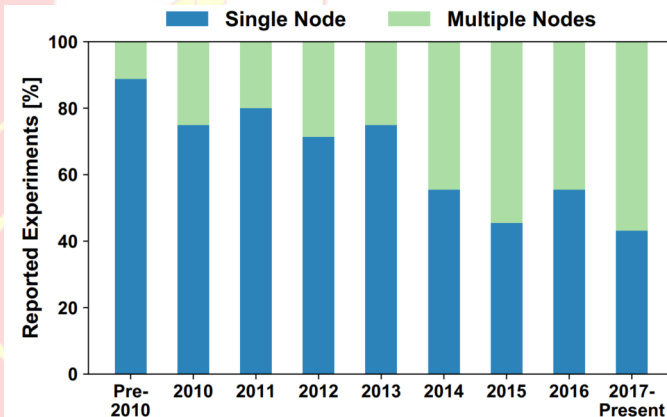
Kiến trúc máy tính song song

Đa máy tính song song (multi-machine Parallelism))

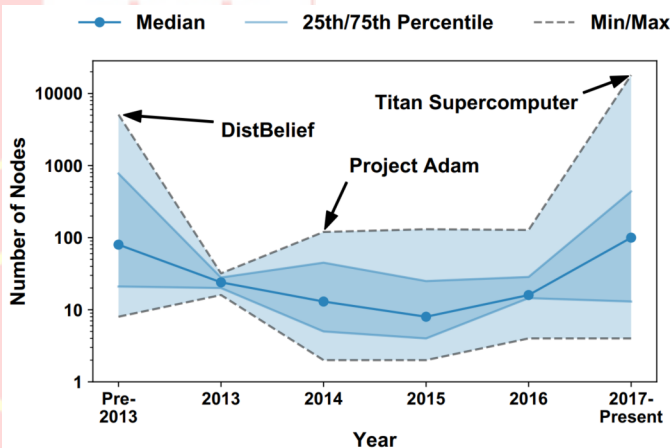
Các chỉ số quan trọng nhất cho mạng kết nối là:

- Độ trễ;
- Băng thông;
- Tỷ lệ truyền tin (Message-Rate)

Kiến trúc máy tính song song



Kiến trúc máy tính song song



Node Count

Kiến trúc máy tính song song

Ngoài ra, kiến trúc song song còn có:

- Lập trình song song (Parallel Programming);
- Thuật toán song song (Parallel Algorithms);



ĐẠI HỌC



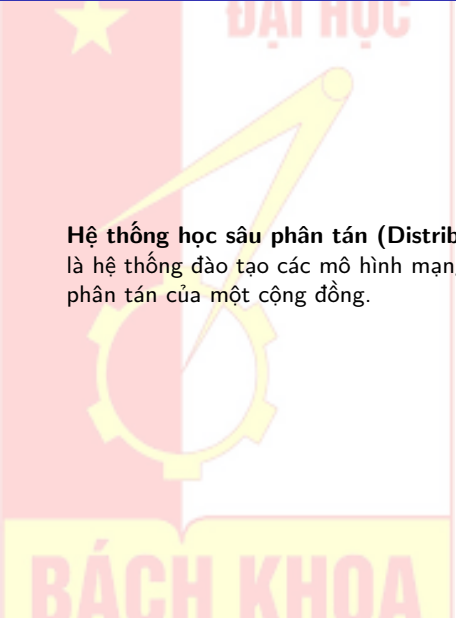
Phần 2: Hệ thống học sâu phân tán

BÁCH KHOA

Hệ thống học sâu phân tán

Hệ thống học sâu phân tán (Distributed Deep Learning Systems - DDLS)

Hệ thống học sâu phân tán



Hệ thống học sâu phân tán (Distributed Deep Learning Systems - DDLS)
là hệ thống đào tạo các mô hình mạng nơ-ron bằng cách sử dụng tài nguyên phân tán của một cộng đồng.

Hệ thống học sâu phân tán

Những vấn đề khi phát triển một hệ thống học sâu phân tán:

Hệ thống học sâu phân tán

Những vấn đề khi phát triển một hệ thống học sâu phân tán:

- ❶ Tính nhất quán;

Hệ thống học sâu phân tán

Những vấn đề khi phát triển một hệ thống học sâu phân tán:

- i** Tính nhất quán;
- ii** Khả năng chịu lỗi;

Hệ thống học sâu phân tán

Những vấn đề khi phát triển một hệ thống học sâu phân tán:

- i Tính nhất quán;
- ii Khả năng chịu lỗi;
- iii Khả năng giao tiếp;

Hệ thống học sâu phân tán

Những vấn đề khi phát triển một hệ thống học sâu phân tán:

- i** Tính nhất quán;
- ii** Khả năng chịu lỗi;
- iii** Khả năng giao tiếp;
- iv** Quản lý tài nguyên;

Hệ thống học sâu phân tán

Những vấn đề khi phát triển một hệ thống học sâu phân tán:

- i** Tính nhất quán;
- ii** Khả năng chịu lỗi;
- iii** Khả năng giao tiếp;
- iv** Quản lý tài nguyên;
- v** Mô hình lập trình.

Hệ thống học sâu phân tán

Các chiến lược song song:

- i Song song dữ liệu (Data Parallelism);
- ii Song song mô hình (Model Parallelism);
- iii Kỹ thuật đường ống (Pipelining);

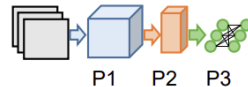
Hệ thống học sâu phân tán



(a) Data Parallelism



(b) Model Parallelism



(c) Layer Pipelining

Hệ thống học sâu phân tán

Song song dữ liệu:

Hệ thống học sâu phân tán

Song song dữ liệu:

- i Chia dữ liệu thành một số phân vùng, với số phân vùng bằng số lượng nút tính toán.
- ii Mỗi nút tính toán đóng vai trò như một công nhân sở hữu một phân vùng độc lập và mỗi công nhân thực hiện tính toán trên phân vùng của chính họ.

Hệ thống học sâu phân tán

Song song dữ liệu

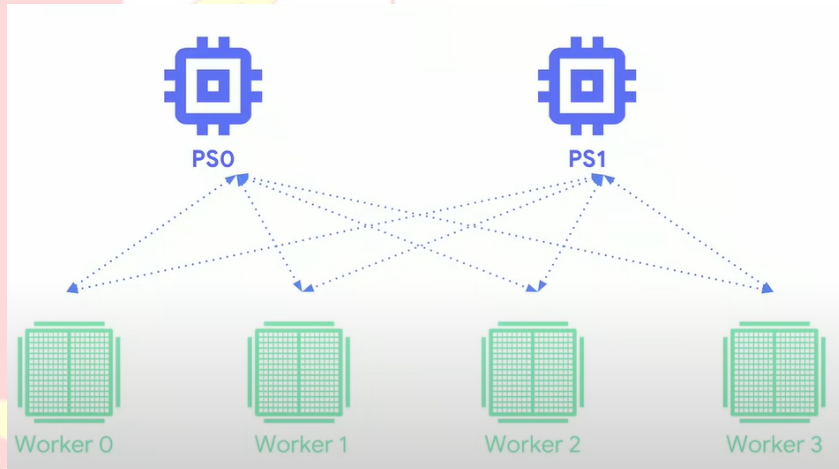
Cách tiếp cận:

- ❶ Máy chủ tham số không đồng bộ (Async Parameter Server);
- ❷ Kiến trúc đồng bộ hóa giảm tất cả (Sync Allreduce Architecture).

Hệ thống học sâu phân tán

Song song dữ liệu

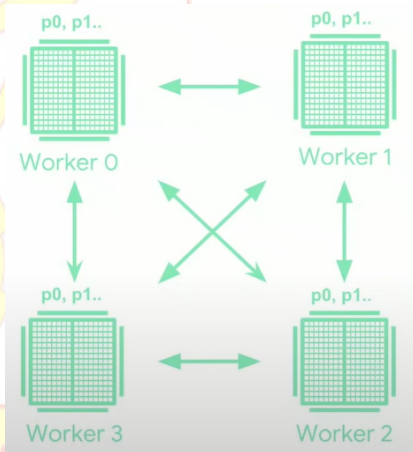
Máy chủ tham số không đồng bộ:



Hệ thống học sâu phân tán

Song song dữ liệu

Kiến trúc đồng bộ hóa giảm tất cả:



Hệ thống học sâu phân tán

Song song mô hình

Thay vì phân vùng dữ liệu như song song dữ liệu, chúng tôi cố gắng phân vùng chính mô hình học sâu để phân phối khối lượng công việc cho nhiều nút (công nhân) tính toán.

Hệ thống học sâu phân tán

Kỹ thuật đường ống

Trong học sâu, pipelining có thể đề cập đến các phép tính chồng chéo, tức là giữa lớp này và lớp tiếp theo; hoặc phân vùng DNN theo độ sâu, gán các lớp cho các bộ xử lý cụ thể.

Một số thuật toán tối ưu hóa phân tán

2 quy trình tối ưu hóa phân tán quy mô lớn:

- i Downpour SGD;
- ii Sandblaster L-BFGS.

Một số thuật toán tối ưu hóa phân tán

Downpour SGD:

Một số thuật toán tối ưu hóa phân tán

Downpour SGD:

- ⓘ Khắc phục hạn chế của SGD;

Một số thuật toán tối ưu hóa phân tán

Downpour SGD:

- i Khắc phục hạn chế của SGD;
- ii Cách tiếp cận cơ bản:

Một số thuật toán tối ưu hóa phân tán

Downpour SGD:

- i Khắc phục hạn chế của SGD;
- ii Cách tiếp cận cơ bản:
 - Chia dữ liệu huấn luyện thành một số tập con;

Một số thuật toán tối ưu hóa phân tán

Downpour SGD:

- i Khắc phục hạn chế của SGD;
- ii Cách tiếp cận cơ bản:
 - Chia dữ liệu huấn luyện thành một số tập con;
 - Chạy một bản sao của mô hình trên các tập con đó;

Một số thuật toán tối ưu hóa phân tán

Downpour SGD:

- i Khắc phục hạn chế của SGD;
- ii Cách tiếp cận cơ bản:
 - Chia dữ liệu huấn luyện thành một số tập con;
 - Chạy một bản sao của mô hình trên các tập con đó;

Một số thuật toán tối ưu hóa phân tán

Sandblaster L-BFGS:

Một số thuật toán tối ưu hóa phân tán

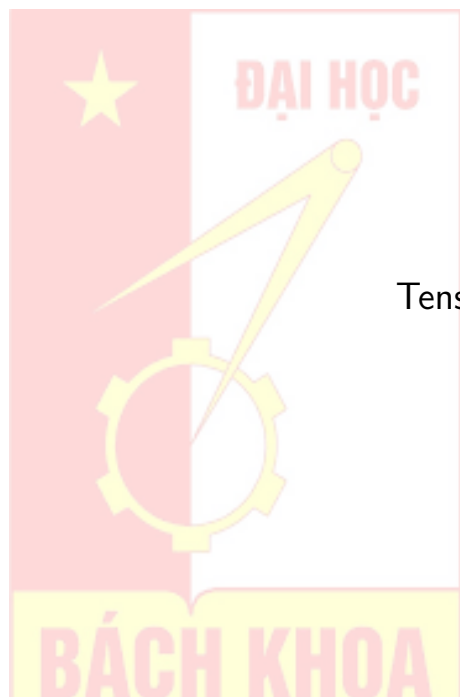
Sandblaster L-BFGS:

- Phương pháp lô (L-BFGS) hoạt động tốt trên mạng học sâu nhỏ;
- Sandblaster giúp cải thiện điều đó;

Một số thuật toán tối ưu hóa phân tán

Sandblaster L-BFGS:

- Phương pháp lô (L-BFGS) hoạt động tốt trên mạng học sâu nhỏ;
- Sandblaster giúp cải thiện điều đó;
- Ý tưởng: lưu trữ và thao tác với tham số phân tán;

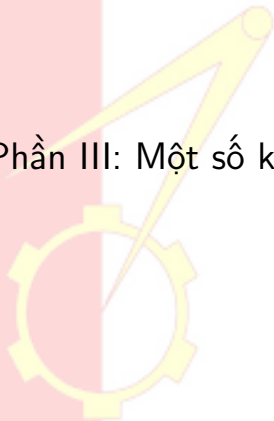


TensorFlow



ĐẠI HỌC

Phần III: Một số kết quả



BÁCH KHOA

Kết quả nghiên cứu tại Google Inc

- Mục đích: đánh giá các thuật toán tối ưu hóa của họ bằng cách áp dụng chúng vào các mô hình đào tạo cho hai vấn đề học sâu khác nhau: nhận dạng đối tượng trong ảnh tĩnh và xử lý âm thanh để nhận dạng giọng nói;

Kết quả nghiên cứu tại Google Inc

- Mục đích: đánh giá các thuật toán tối ưu hóa của họ bằng cách áp dụng chúng vào các mô hình đào tạo cho hai vấn đề học sâu khác nhau: nhận dạng đối tượng trong ảnh tĩnh và xử lý âm thanh để nhận dạng giọng nói;
- Nhận dạng giọng nói: phân loại trạng thái âm thanh:

Kết quả nghiên cứu tại Google Inc

- Mục đích: đánh giá các thuật toán tối ưu hóa của họ bằng cách áp dụng chúng vào các mô hình đào tạo cho hai vấn đề học sâu khác nhau: nhận dạng đối tượng trong ảnh tĩnh và xử lý âm thanh để nhận dạng giọng nói;
- Nhận dạng giọng nói: phân loại trạng thái âm thanh:
 - ⊕ Sử dụng mô hình học sâu 5 lớp, trong đó:
 - Lớp ẩn: 2560 nút;
 - Lớp đầu ra: 8192 nút;
 - ⊕ 42 triệu tham số;
 - ⊕ 1,1 tỷ bản ghi.

Kết quả nghiên cứu tại Google Inc

- Mục đích: đánh giá các thuật toán tối ưu hóa của họ bằng cách áp dụng chúng vào các mô hình đào tạo cho hai vấn đề học sâu khác nhau: nhận dạng đối tượng trong ảnh tĩnh và xử lý âm thanh để nhận dạng giọng nói;
- Nhận dạng giọng nói: phân loại trạng thái âm thanh:
 - ⊕ Sử dụng mô hình học sâu 5 lớp, trong đó:
 - Lớp ẩn: 2560 nút;
 - Lớp đầu ra: 8192 nút;
 - ⊕ 42 triệu tham số;
 - ⊕ 1,1 tỷ bản ghi.
- Xử lý ảnh:

Kết quả nghiên cứu tại Google Inc

- Mục đích: đánh giá các thuật toán tối ưu hóa của họ bằng cách áp dụng chúng vào các mô hình đào tạo cho hai vấn đề học sâu khác nhau: nhận dạng đối tượng trong ảnh tĩnh và xử lý âm thanh để nhận dạng giọng nói;
- Nhận dạng giọng nói: phân loại trạng thái âm thanh:
 - ⊕ Sử dụng mô hình học sâu 5 lớp, trong đó:
 - Lớp ẩn: 2560 nút;
 - Lớp đầu ra: 8192 nút;
 - ⊕ 42 triệu tham số;
 - ⊕ 1,1 tỷ bản ghi.
- Xử lý ảnh:
 - ⊕ 16 triệu ảnh;
 - ⊕ kích thước mỗi ảnh: 100x100 pixel;
 - ⊕ 3 giai đoạn: lọc, gộp và chuẩn hóa tương phản cục bộ;

Kết quả nghiên cứu tại Google Inc



Mô hình gọng nôi:

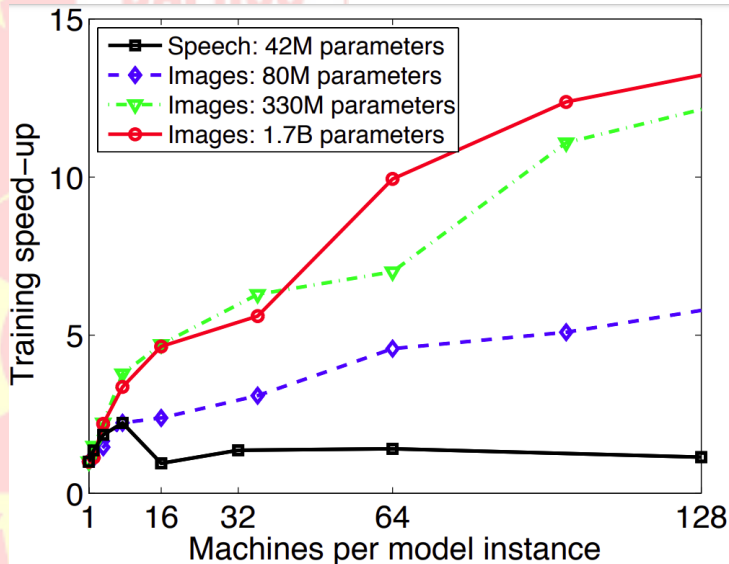
Kết quả nghiên cứu tại Google Inc

- Mô hình gong nói: Chạy trên 8 máy, có tốc độ tính toán **nh nhanh hơn 2,2 lần** so với sử dụng một máy duy nhất.

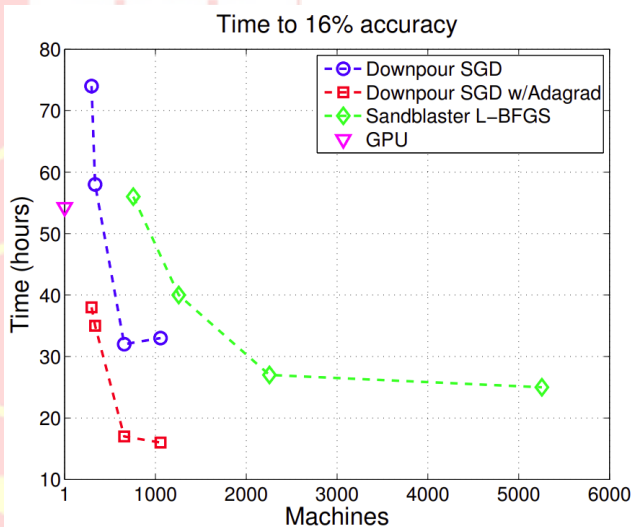
Kết quả nghiên cứu tại Google Inc

- Mô hình gong nói: Chạy trên 8 máy, có tốc độ tính toán **nhANH hơn 2,2 lần** so với sử dụng một máy duy nhất.
- Mô hình xử lý ảnh: Tốc độ nhanh hơn 12 lần khi sử dụng 81 máy.

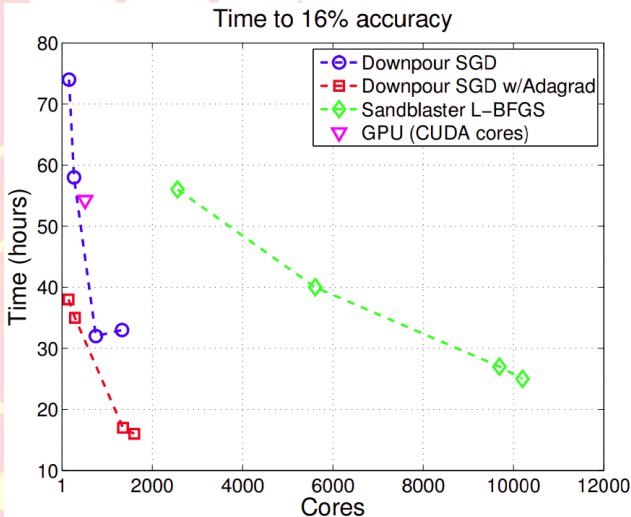
Kết quả nghiên cứu tại Google Inc



Kết quả nghiên cứu tại Google Inc



Kết quả nghiên cứu tại Google Inc



Kết quả nghiên cứu tại MIT

- Mục đích: Dự báo lượng mưa;

Kết quả nghiên cứu tại MIT

- Mục đích: Dự báo lượng mưa;
- Triển khai trong TensorFlow 1.12 bằng cách sử dụng API Keras. Mạng này có 17.395.992 tham số có thể huấn luyện.

Kết quả nghiên cứu tại MIT

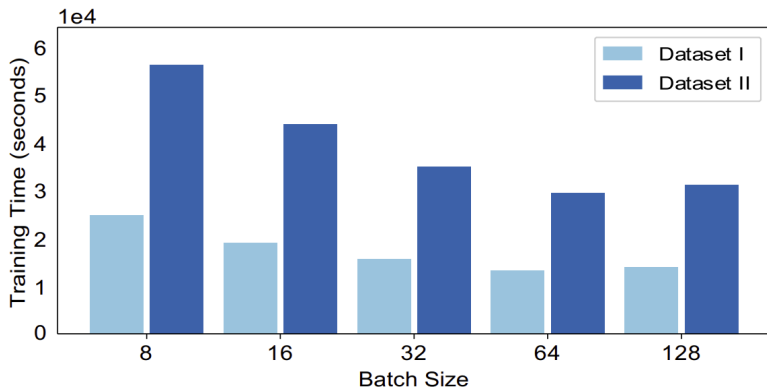
- Mục đích: Dự báo lượng mưa;
- Triển khai trong TensorFlow 1.12 bằng cách sử dụng API Keras. Mạng này có 17.395.992 tham số có thể huấn luyện.
- Chạy bộ dữ liệu đã có trên 1 GPU GK210, thu được:

	Số lượng ảnh đào tạo	Số lượng ảnh kiểm tra	Epochs	Thời gian đào tạo (giờ)
Bộ DL 1	17,833	10,052	100	23.219
Bộ DL 2	45,897	10,052	100	59.136

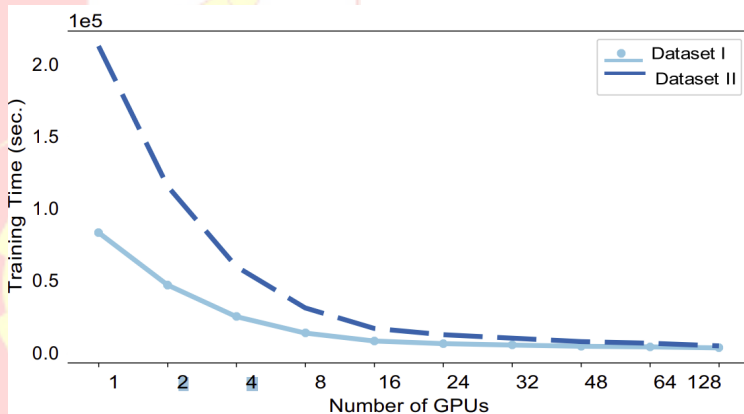
Kết quả nghiên cứu tại MIT

- TensorFlow / Keras và sử dụng framework Horovod;
- 32 nút với 4 thiết bị GK210 trên mỗi nút, trong tổng số 128 thiết bị GPU.

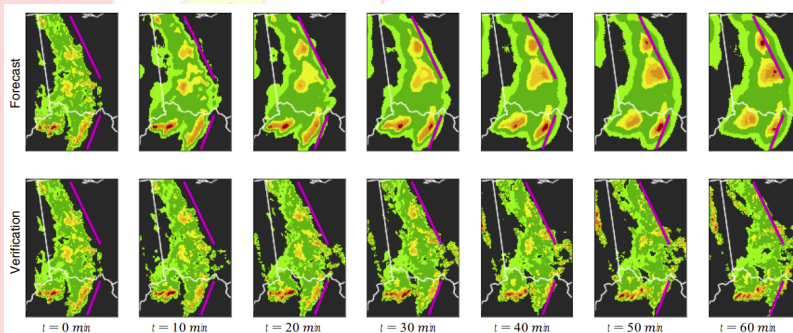
Kết quả nghiên cứu tại MIT



Kết quả nghiên cứu tại MIT



Kết quả nghiên cứu tại MIT



Kết quả chương trình Demo

- Thực hiện trên Ubuntu với thư viện phân tán của TensorFlow;


Kết quả chương trình Demo

- Thực hiện trên Ubuntu với thư viện phân tán của TensorFlow;
- Thực hiện trên 1 thiết bị laptop và mở nhiều cổng;
- Các cổng thông nhau thông qua giao thức TCP/IP.

Listing 6: Hàm tạo ra các tiến trình (processes)

```
import subprocess

subprocess.Popen('python3_asyn_distributed_tf.py --job_name="ps" --task_index_0',
shell = True)
subprocess.Popen('python3_asyn_distributed_tf.py --job_name="worker" --task_index_0',
shell = True)
subprocess.Popen('python3_asyn_distributed_tf.py --job_name="worker" --task_index_1',
shell = True)
subprocess.Popen('python3_asyn_distributed_tf.py --job_name="worker" --task_index_2',
shell = True)
```



Listing 7: Khởi tạo các cổng

```
parameter_servers = ["localhost:2222"]  
workers = [ "localhost:2223", "localhost:2224", 'localhost:_2225 ']  
cluster = tf.train.ClusterSpec({"ps":parameter_servers , "worker":workers})
```



ĐẠI HỌC



python3	11308	huuthuat	5u	IPv6	190768	0t0	TCP	*:2222	(LISTEN)
python3	11310	huuthuat	5u	IPv6	182689	0t0	TCP	*:2223	(LISTEN)
python3	11312	huuthuat	5u	IPv6	180875	0t0	TCP	*:2224	(LISTEN)
python3	11313	huuthuat	5u	IPv6	192589	0t0	TCP	*:2225	(LISTEN)



BÁCH KHOA





ĐẠI HỌC

```
Test accuracy at step 50: 0.826
Test accuracy at step 60: 0.815
Test accuracy at step 70: 0.827
Test accuracy at step 80: 0.823
Test accuracy at step 90: 0.823
Worker 0, At step 468, Cost: 0.3539 Accuracy: 0.8906 AvgTime: 0.18ms
Worker 0, At step 568, Cost: 0.3790 Accuracy: 0.8594 AvgTime: 18.17ms
Worker 0, At step 668, Cost: 0.3993 Accuracy: 0.8594 AvgTime: 17.98ms
Worker 0, At step 768, Cost: 0.3399 Accuracy: 0.8672 AvgTime: 17.65ms
Worker 0, At step 868, Cost: 0.3623 Accuracy: 0.8203 AvgTime: 18.30ms
Test accuracy at step 0: 0.846
Test accuracy at step 10: 0.848
Test accuracy at step 20: 0.844
Test accuracy at step 30: 0.853
Test accuracy at step 40: 0.848
Test accuracy at step 50: 0.836
Test accuracy at step 60: 0.845
Test accuracy at step 70: 0.843
Test accuracy at step 80: 0.843
Test accuracy at step 90: 0.841
Worker 0, At step 936, Cost: 0.2671 Accuracy: 0.8906 AvgTime: 0.21ms
Done!!!!
```

BÁCH KHOA





ĐẠI HỌC

```
Test accuracy at step 50: 0.826
Test accuracy at step 60: 0.815
Test accuracy at step 70: 0.827
Test accuracy at step 80: 0.823
Test accuracy at step 90: 0.823
Worker 0, At step 468, Cost: 0.3539 Accuracy: 0.8906 AvgTime: 0.18ms
Worker 0, At step 568, Cost: 0.3790 Accuracy: 0.8594 AvgTime: 18.17ms
Worker 0, At step 668, Cost: 0.3993 Accuracy: 0.8594 AvgTime: 17.98ms
Worker 0, At step 768, Cost: 0.3399 Accuracy: 0.8672 AvgTime: 17.65ms
Worker 0, At step 868, Cost: 0.3623 Accuracy: 0.8203 AvgTime: 18.30ms
Test accuracy at step 0: 0.846
Test accuracy at step 10: 0.848
Test accuracy at step 20: 0.844
Test accuracy at step 30: 0.853
Test accuracy at step 40: 0.848
Test accuracy at step 50: 0.836
Test accuracy at step 60: 0.845
Test accuracy at step 70: 0.843
Test accuracy at step 80: 0.843
Test accuracy at step 90: 0.841
Worker 0, At step 936, Cost: 0.2671 Accuracy: 0.8906 AvgTime: 0.21ms
Done!!!!
```

BÁCH KHOA



