RESEARCH PROPOSAL

Song song hóa thuật toán Random Forest dùng cho việc phân loại

Group name: 3rd-team

List of members:

19127496 - Trương Quang Minh Nhật

19127503 - Ngô Quốc Phát

19127347 - Kiều Hải Đăng

Keywords: Song song, randomforest, decisiontree, cuda, python, numba.

List of references:

- github.com/harrypnh/random-forest-from-scratch.

Content

1. **Summary:** Mục tiêu chính của nhóm là cải thiện tốc độ thuật toán Random Forest có sẵn, càng nhanh càng tốt. Nhóm sẽ có tất cả 4 mục tiêu để so sánh: chạy tuần tự, chạy song song bằng CPU, chạy song song bằng GPU, sử dụng shared memory để chạy song song.
2. **Background**

Mục đích chính cho tới thời điểm viết proposal này của nhóm là cải thiện tốc độ thuật toán Random Forest nói trên bằng cách sử dụng numba. Cụ thể nhóm sẽ xây dựng ra tất cả 4 version:

- V1 - là thuật toán Random Forest chạy tuần tự.
- V2 - là thuật toán Random Forest chạy song song bằng CPU.
- V3 - là thuật toán Random Forest chạy song song bằng GPU.
- V4 - là thuật toán Random Forest chạy song song sử dụng shared memory.

Các version sau sẽ được xây dựng nếu xong V4 và nếu vẫn còn có thể cải thiện tốc độ.

3. **The challenge**

Vấn đề khó khăn nhất là các thành viên chưa từng tiếp xúc với numba và lập trình song song nên kiến thức và kinh nghiệm còn rất hạn chế. Việc đọc, hiểu code trên mạng cũng mất rất nhiều thời gian.

Bên cạnh đó thì tài liệu về nội dung tìm hiểu cũng khá hạn chế.

4. Resources

Nhóm tìm được thuật toán Random Forest từ một nguồn trên github, họ implement thuật toán với 2 mục đích chính: chẩn đoán bệnh ung thư vú và đánh giá ô tô (Car Evalution).

Tại sao lại chọn từ nguồn này?

Vì họ xây dựng thuật toán Random Forest theo hàm chứ không xây theo class, điều này khiến cho quá trình làm việc với numba trở nên dễ dàng hơn.

Về phần dataset, có lẽ nhóm sẽ tìm một dataset nào đó đủ lớn và dễ hiểu để training model trước. Việc này sẽ dễ hơn là dùng lại dataset có sẵn vì các kiến thức về xe ô tô và y tế nhóm chưa có nhiều kinh nghiệm nên việc diễn giải sẽ khó hơn.

5. Goals And Deliverables

Describe the deliverables or goals of your project. This is by far the most important section of the proposal!

- Ở mức kỳ vọng 100% - nhóm mong muốn có thể xây dựng được thuật toán đến V3 (GPU) và tốc độ có sự cải thiện khi chạy với dataset đủ lớn.
- Ở mức kỳ vọng 125% - nhóm mong muốn xây dựng được thuật toán đến V4 và có thể ứng dụng thuật toán của nhóm vào để thử giải quyết một bài toán thực tế nào đó do các thành viên nhóm đề cử (ví dụ chuẩn đoán bệnh).
- Kế hoạch cho buổi seminar - trước tiên cần các thành viên hiểu về những thứ nhóm đang làm và hiểu về công việc của mình cũng như các thành viên khác. Về phần trình bày, nhóm dự kiến sẽ trình bày theo thứ tự lần lượt từ V1 - V4. Đối với V1, nhóm sẽ trình bày cách mình hiểu về thuật toán và những chỗ có thể triển khai

song song được. Đối với các version sau, nhóm sẽ giới thiệu về cách nhóm chạy song song, lý do tại sao chạy như vậy nếu có.

- Nhóm kỳ vọng tốc độ sẽ tăng 10x lần cho version chạy song song đầu tiên, và sẽ tăng liên tục cho các version tiếp nối sau đó.

Weekly schedule:

<table>
<tr><th></th><th>Week 01 (from 13-06 to 20-06)</th><th>Week 02 (from 21-06 to 27-06)</th><th>Week 03 (from 28-06 to 04-07)</th><th>Week 04 (from 05-07 to 11-07)</th><th>Week 05 (from 12-07 to 18-07)</th><th>Week 06 (from 19-07 to 25-07)</th><th>Week 07 (from 26-07 to 25-07)</th></tr>
<tr><td>Đăng</td><td rowspan="3">Tìm kiếm nguồn tài liệu, thuật toán, code, …</td><td>Chạy thuật toán V1</td><td>Chạy thuật toán V2</td><td rowspan="2">Chạy thuật toán V3</td><td>Viết báo cáo</td><td rowspan="3">Tiếp tục chạy V4 nếu cần thiết. Nếu xong rồi thì có thể tìm ý tưởng thực tế nào đó có thể giải quyết bằng model và thử áp dụng.</td><td rowspan="3">Hoàn thành tiến độ, tổng kết lại mọi thứ để chuẩn bị cho báo cáo cuối kỳ.</td></tr>
<tr><td>Phát</td><td>Đọc hiểu thuật toán (do cần thêm thời gian)</td><td>Nghiên cứu ý tưởng cho V3</td><td rowspan="2">Chạy thuật toán V4</td></tr>
<tr><td>Nhật</td><td>Tìm kiếm những chỗ có thể chạy song song được.</td><td>Nghiên cứu ý tưởng cho V4</td><td>Báo cáo kết quả nghiên cứu và cách thực hiện, chạy thử V4</td></tr>
</table>