NGHIÊN CỨU THỬ NGHIỆM KẾT HỢP MÔI TRƯỜNG LÀM VIỆC GOOGLE COLABORATORY VÀ PHƯƠNG PHÁP HỌC MÁY (MACHINE LEARNING) TRONG PHÂN LOẠI ẢNH VIỄN THÁM

PHẠM MINH HẢI[1], NGUYỄN NGỌC QUANG[2]

[1]Viện Khoa học Đo đạc và Bản đồ, [2]Cục Viễn thám Quốc gia

Tóm tắt:

Khi độ chính xác và mật độ dữ liệu tăng theo thời gian, khối lượng dữ liệu tăng theo cấp số nhân nên đó thực sự là nguồn dữ liệu vô cùng khổng lồ mà trong những năm gần đây có dùng thuật ngữ "Big data" để mô tả . Dữ liệu viễn thám là dữ liệu có cấu trúc phức tạp, nhiều định dạng do đó cần phải thiết kế hệ thống có kiến trúc lưu trữ loại big data viễn thám này. Bên cạnh đó, một môi trường có thể xử lý với tốc độ nhanh, khả năng ứng dụng được các phương pháp học máy để xử lý dữ liệu viễn thám cần được quan tâm nghiên cứu phát triển. Trong phạm vi bài báo này, nhóm nghiên cứu đã thử nghiệm tính khả dụng của môi trường Google Colaboratory phục vụ phân loại ảnh viễn thám.

1. Giới thiệu chung

Trong những năm gần đây, sự phát triển công nghệ số đã khởi xướng việc phổ biến dữ liệu viễn thám rộng rãi tới người sử dụng. Hiện tại, đã có hơn 1000 vệ tinh viễn thám đã được phóng lên quỹ đạo [1] và dữ liệu thu được tại trạm thu ảnh vệ tinh được lưu trữ nhiềuTerabyte mỗi ngày [2]. Theo thống kê của Hệ thống thông tin và dữ liệu hệ thống quan sát trái đất (EOSDIS) năm 2014, EOSDIS quản lý hơn 9 Petabyte dữ liệu và hàng ngày nhận thêm 6.4 Terabyte vào kho lưu trữ (NASA 2016). Ở cơ quan vũ trụ châu Âu lượng dữ liệu ảnh viễn thám thu nhận được đã vượt quá 1.5 Petabyte [3], còn nếu xét tổng dung lượng dữ liệu viễn thám đã thu nhận được thì đã đạt tới đơn vị Zetabyte (10e9 Terabyte) [4].

Khi độ chính xác và mật độ dữ liệu tăng theo thời gian, khối lượng dữ liệu tăng theo cấp số nhân nên đó thực sự là nguồn dữ liệu vô cùng khổng lồ mà trong những năm gần đây có dùng thuật ngữ "Big data" để mô tả [5] mà giờ thực tế đã công nhận dữ liệu viễn thám là 'big data'. Thêm vào đó, dữ liệu viễn thám có cấu trúc rất phức tạp, nhiều định dạng như Geotiff, ASCII, HDF,…và không có sự tương tác giữa các loại dữ liệu từ các vệ tinh viễn thám khác nhau do đó cần phải thiết kế hệ thống có kiến trúc lưu trữ loại big data viễn thám này. Một vấn đề nữa là xử lý dữ liệu viễn thám đặt ra yêu cầu cao về hiệu năng tính toán. Một mặt, với sự cải tiến liên tục về chất lượng và độ chính xác của dữ liệu, dữ liệu có độ phân giải cao hơn cần được xử lý; mặt khác, với sự phát triển của các thuật toán như machine learning và deep learning, các thuật toán xử lý dữ liệu viễn thám ngày càng trở nên phức tạp.

Để giải quyết các vấn đề trên, các nhà khoa học đã nỗ lực tập trung vào tính khả dụng của dữ liệu viễn thám và khả năng xử lý. Để đảm bảo tính sẵn sàng ở mức độ cao của dữ liệu viễn thám, các hệ thống lưu trữ phân tán đã được áp dụng rộng rãi. Tiêu biểu như MongeDB, một cơ sở dữ liệu phân tán ban đầu hỗ trợ cả lưu trữ và lập chỉ mục dữ liệu viễn thám và dữ liệu vectơ [6,7]. Hệ thống tệp phân tán Hadoop (HDFS) được áp dụng để có thể lưu trữ tất cả các loại dữ liệu viễn thám, nó đã chứng tỏ là vượt trội so với hệ thống tệp cục bộ [8,9]. Với cơ sở dữ liệu NoQuery cũng có thể lưu trữ dữ liệu viễn thám như HBase. Ngoài ra, các hệ thống lưới toàn cầu riêng biệt (DGGS) và một số cách tiếp cận tổ chức dữ liệu khác cũng giúp lập chỉ mục và xác định tổ chức dữ liệu hang. HPC dựa trên cluster và cloud là hai kiểu chiếm ưu thế nhất để xử lý

Ngày nhận bài: 07/02/2020, ngày chuyển phản biện: 12/02/2020, ngày chấp nhận phản biện: 17/02/2020, ngày chấp nhận đăng: 22/02/2020

viễn thám. Cấu trúc của Master-Slave giúp lập kế hoạch và thực hiện xử lý viễn thám phức tạp, điều này chứng tỏ cải thiện đáng kể hiệu quả của tính toán trong xử lý dữ liệu viễn thám. OpenMP cung cấp hiệu suất tính toán linh hoạt, có thể mở rộng và có khả năng tính toán.

Ngoài các giải pháp riêng lẻ, một số nền tảng hợp nhất được đề xuất để cung cấp giải pháp xuyên suốt cho viễn thám dữ liệu lớn. Google Earth Engine (GEE) là một cái tên không còn xa lạ đặc biệt với người sử dụng cá nhân-còn nhiều hạn chế về hạ tầng lưu trữ và tính toán hiệu năng cao, GEE cung cấp quyền truy cập dễ dàng để sử dụng các tài nguyên tính toán dựa vào nền tảng cloud-computing cho các bộ dữ liệu viễn thám quy mô lớn. Tuy nhiên, GEE không phải là nguồn mở và không thuận tiện khi xử lý các bộ dữ liệu riêng với tài nguyên máy tính riêng của người dung mặc dù đây là một nền tảng xử lý dữ liệu lớn rất thành công.

Vì thế Google Colaboratory (GC) ra đời để hoàn thành nốt sứ mạng trên, rất phù hợp để giải các bài toán đòi hỏi hiệu năng tính toán lớn, tích hợp sẵn các framework như Tensorflow, Keras và PyTorch để hỗ trợ cho deep learning và đặc biệt là hoàn toàn miễn phí cho người sử dụng, đáp ứng được nhu cầu trong lĩnh vực nghiên cứu và giáo dục mà không phải chọn giải pháp thuê dịch vụ của Amazon Web Services (AWS) như trước kia. Trong phạm vi bài báo này, nhóm nghiên cứu đã thử nghiệm tính khả dụng của ừn dụng môi trường GC kết hợp với phương pháp học máy trong phân loại ảnh viễn thám.

2. Khái quát về Google Colaboratory

Google đã rất tích cực trong nghiên cứu về Trí tuệ nhân tạo (AI), trong nhiều năm Google đã phát triển một nền tảng AI gọi là TensorFlow và công cụ Colaboratory. Colaboratory hay gọi đơn giản Colab, cung cấp dịch vụ cloud-computing miễn phí sử dụng môi trường Jupyter notebook nên không yêu cầu phải cài đặt để sử dụng, cùng với Google docs nhiều người sử dụng có thể cộng tác cùng làm việc một lúc trên một chương trình.

Hiện tại, Colab cung cấp dịch vụ điện toán sử dụng GPU (Tesla K80) và TPU (TPUv2) tối đa 12 giờ cho mỗi phiên làm việc, sau 12 giờ Colab sẽ chỉ định một máy ảo khác để phục vụ, chu kỳ cứ thế lặp lại và không có giới hạn cho việc có bao nhiêu máy ảo có thể được sử dụng thông quan một tài khoản. Một điểm cần thiết phải lưu ý là sau thời gian phiên làm việc là 12 giờ, người sử dụng sẽ bị mất quyền truy cập vào máy ảo đã được chỉ định đó khi đó tất cả các bộ dữ liệu, các tham số mô hình sẽ không được lưu vào ổ Google, vì vậy hãy đảm bảo việc sao lưu quick-look cũng như các tham số mô hình theo định kỳ, nếu không sẽ phải training lại từ đầu.

Google Colab được khuyến cáo khi thử nghiệm môi trường GC với cấu hình máy tính:

- Processor: Intel Xeon 2.3GHz (04 Processor, 1 Processor: 02 Cores)

- Ram memory: 25Mb

- Memory: 34 Gb

- Graphic: GPU

(Xem bảng 1)

Tuy nhiên, nhóm thực hiện nghiên cứu đã thử nghiệm ở cấu hình thấp hơn để kiểm tra tính khả thi khi vận hành xử lý dữ liệu ở môi trường máy tính có tốc độ trung bình:

Bảng 1: Các thông số kỹ thuật khác của cấu hình máy tính được giới thiệu khi sử dụng GC

```
+-----------------------------------------------------------------------------+
| NVIDIA-SMI 440.48.02    Driver Version: 418.67       CUDA Version: 10.1     |
|-------------------------------+----------------------+----------------------+
| GPU  Name        Persistence-M| Bus-Id        Disp.A | Volatile Uncorr. ECC |
| Fan  Temp  Perf  Pwr:Usage/Cap|         Memory-Usage | GPU-Util  Compute M. |
|===============================+======================+======================|
|   0  Tesla P100-PCIE...  Off  | 00000000:00:04.0 Off |                    0 |
| N/A   31C    P0    25W / 250W |      0MiB / 16280MiB |      0%      Default |
+-------------------------------+----------------------+----------------------+
```

- Processor 2.7 GHz Core i5 (1 Processor 02 Cores)

- Ram memory: 8Gb DDR3 1867 Mhz,

- Graphic: Intel Iris Graphics 6100 1536 Mb

3. Thử nghiệm với phương pháp phân loại Kmeans Classification và Random Foresr

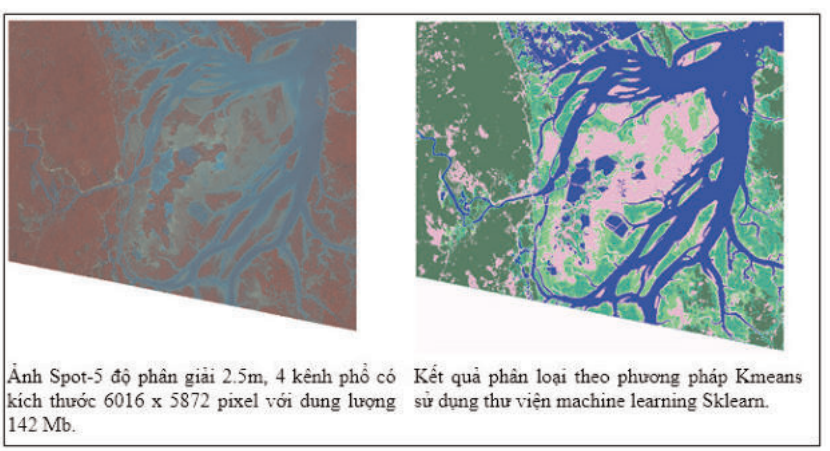
Dưới đây là thử nghiệm ứng dụng Google Colab (GC) trong xử lý dữ liệu viễn thám cụ thể là phân loại dữ liệu viễn thám để thấy được sức mạnh tính toán của GC so với máy tính thông thường. Phương pháp phân loại sử dụng bộ thư viện machine learning Sklearn hỗ trợ cho ngôn ngữ lập trình Python. Phân loại ảnh viễn thám là một công đoạn xử lý mất rất nhiều công và thời gian tính toán bằng máy tính thông thường tốn từ vài giờ đến hàng chục giờ, đặc biệt với phương pháp phân loại có giám sát đòi hỏi phải training một lượng lớn mẫu trên ảnh.

Phương pháp thực hiện tiến hành phân loại ảnh cho cả phương pháp phân loại không giám sát theo phương pháp Kmeans Classification và phương pháp phân loại có giám sát tương đối ưu việt hiện nay là Random Forest Classification.

a. Kết quả thử nghiệm với phương pháp phân loại Kmeans

Nhóm nghiên cứu sử dụng ảnh Spot-5 độ phân giải 2.5m, 4 kênh phổ có kích thước 6016 x 5872 pixel với dung lượng 142 Mb. Phương pháp phân loại Kmeans được thử nghiệm dựa vào thư viện Sklearn viết trên ngôn ngữ lập trình Python 3.7, với số lớp cần phân loại là 10, vòng lặp tính ít nhất là 10 và tối đa là 300, thuật toán Kmeans bao gồm 'auto', 'full' và 'elkan', thử nghiệm này đặt 'auto'. Kết quả phân loại Kmeans cho thấy thời gian chạy trên ứng dụng lập trình xử lý trên máy tính là 208.13 phút, còn thời gian chạy trên GC là 15.23 phút. *(Xem hình 1, 2)*

Hình 1: So sánh thời gian tính toán giữa hai môi trường trên máy tính cá nhân



Hình 2: Kết quả phân loại ảnh viễn thám sử dụng thư viện machine learning Sklearn

b. Kết quả thử nghiệm với phương pháp phân loại Random Forest Classification

Nhóm nghiên cứu sử dụng dữ liệu thử nghiệm là ảnh Spot-6 độ phân giải 1.5m, 3 kênh phổ có kích thước 9375 x 8989 với dung lượng 506 Mb. Bảng giải đoán (training data) gồm 10 mẫu. Trong đó mẫu ít nhất gồm 1352 pixels, mẫu nhiều nhất là 83607119 pixels.

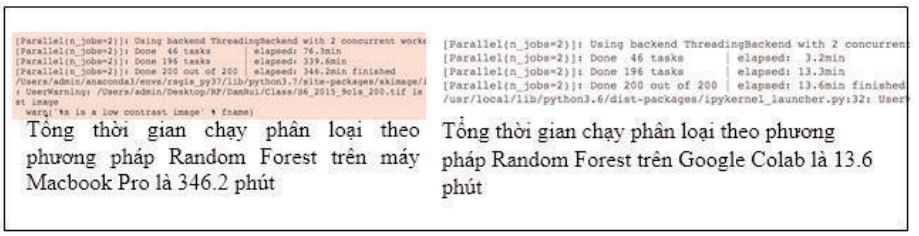
Thuật toán Random Forest Classification được thử nghiệm dựa vào thư viện Sklearn viết trên ngôn ngữ lập trình Python 3.7. Tham số đầu vào của thử nghiệm gồm: số trees là 200, kiểu 'Gini', mẫu nhỏ nhất có thể chia: 2, số leaf nhỏ nhất:1, n_jobs: 2.

Kết quả thử nghiệm này cho thấy máy tính xử lý hết 346.2 phút còn GC chỉ mất 13.6 phút để hoàn thành. *(Xem hình 3, 4)*

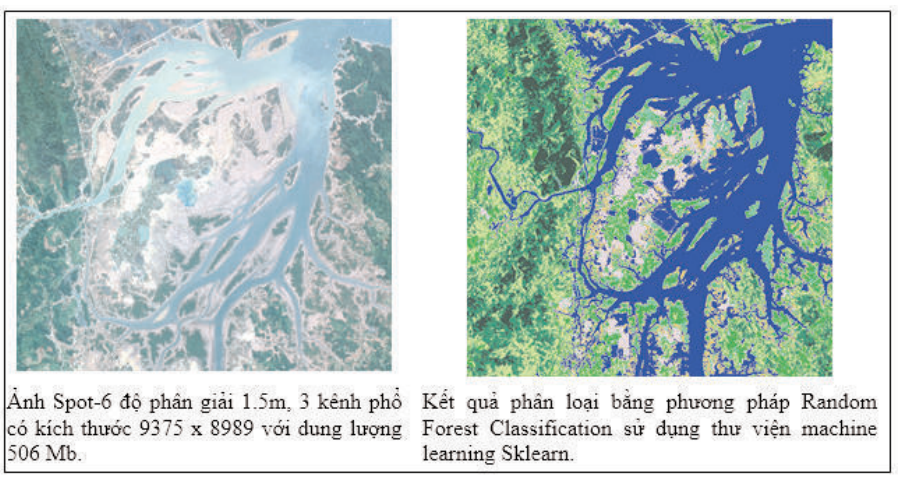
4. Kết luận

Kết quả thử nghiệm đã minh chứng lợi thế về thời gian tính toán của GC, đặc biệt với các bài toán sử dụng dữ liệu có dung lượng lớn thì GC càng tỏ rõ sự vượt trội với tốc độ tính toán gấp đến 25 lần so với máy tính thông thường với thử nghiệm phân loại Random Forest Classification. CG vẫn phát huy được hiệu suất xử lý dữ liệu nhanh khi sử dụng với máy tính cá nhân thông thường. Với ứng dụng trí tuệ nhân tạo (AI) ngày càng phát triển chóng mặt đặc biệt trong lĩnh vực ứng dụng từ ảnh viễn thám thì GC đã mang tới một cơ hội lớn cho người sử dụng, đó là một công cụ hỗ trợ cực kỳ quan trọng mang tính quyết định mà không phải trả bất kỳ khoản phí sử dụng nào.

GC cũng có thể xem xét như là một nguồn cung cấp hạ tầng tính toán hiệu năng cao, kèm theo một hạ tầng lưu trữ thông qua Google Drive với mức phí hợp lý (nếu vượt quá dung lượng cấp miễn phí của Google) cho các cơ quan nghiên cứu và ứng dụng nhằm cắt giảm hẳn những chi phí không cần thiết để trang bị hệ



Hình 3: So sánh thời gian tính toán giữa hai môi trường trên máy tính cá nhân



Hình 4: Kết quả phân loại ảnh viễn thám sử dụng thư viện Random Forest

thống phần cứng, phần mềm kèm kinh phí duy trì hoạt động và bảo trì bảo dưỡng hệ thống định kỳ.❍

Tài liệu tham khảo

[1]. Drahansky, M.; Paridah, M.; Moradbak, A.; Mohamed, A.; Owolabi, F.; Abdulwahab, T.; Asniza, M.; Abdul, K.S.H. A Review: Remote Sensing Sensors. *IntechOpen* 2016, *17*, 777. [Google Scholar]

[2]. Gamba, P.; Du, P.; Juergens, C.; Maktav, D. Foreword to the Special Issue on Human Settlements: A Global Remote Sensing Challenge. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* 2011, *4*, 5–7. [Google Scholar] [CrossRef]

[3]. He, G.; Wang, L.; Ma, Y.; Zhang, Z.; Wang, G.; Peng, Y.; Long, T.; Zhang, X. Processing of earth observation big data: Challenges and countermeasures. *Kexue Tongbao Chin. Sci. Bull.* 2015, *60*, 470–478. [Google Scholar]

[4]. Guo, H.; Wang, L.; Chen, F.; Liang, D. Scientific big data and Digital Earth. *Chin. Sci. Bull.* 2014, *59*, 5066–5073. [Google Scholar] [CrossRef]

[5]. Chang WL, Grady N (2015) NIST big data interoperability framework: volume 1, big data definitions (No. special publication (NIST SP)-1500-1).

[6]. Huang, B.; Jin, L.; Lu, Z.; Yan, M.; Wu, J.; Hung, P.C.K.; Tang, Q. RDMA-driven MongoDB: An approach of RDMA enhanced NoSQL paradigm for large-Scale data processing. *Inf. Sci.* 2019, *502*, 376–393. [Google Scholar] [CrossRef]

[7]. Li, C.; Yang, W. The distributed storage strategy research of remote sensing image based on Mongo DB. In Proceedings of the 2014 3rd International Workshop on Earth Observation and Remote Sensing Applications (EORSA), Changsha, China, 11–14 June 2014; pp. 101–104. [Google Scholar]

[8]. Liu, X.; Han, J.; Zhong, Y.; Han, C.; He, X. Implementing WebGIS on Hadoop: A case study of improving small file I/O performance on HDFS. In Proceedings of the 2009 IEEE International Conference on Cluster Computing and Workshops, New Orleans, Louisiana, 31 August–4 September 2009; pp. 1–8. [Google Scholar]

[9]. Lin, F.C.; Chung, L.K.; Ku, W.Y.; Chu, L.R.; Chou, T.Y. The framework of cloud computing platform for massive remote sensing images. In Proceedings of the 2013 IEEE 27th International Conference on Advanced Information Networking and Applications (AINA), Barcelona, Spain, 25–28 March 2013; pp. 621–628. [Google Scholar].❍

Summary

Using Google Colaboratory with machine learning for the satellite image classification

Pham Minh Hai, Vietnam Institute of Geodesy and Cartography

Nguyen Ngoc Quang, National Remote Sensing Department

As an increase in the accuracy and types of data, a thousand of Terabytes of spatial data has become an huge data source recently called "Big data". Remote sensing data has complex data structure and many formats, so that it is necessary to develop an environment for the purposes of data processing and data storage. In addition, that environment can process satellite data fast and stably, and be able to apply machine learning methods for processing remote sensing data. In this manuscript, we will take an investigation in using Google Colaboratory with machine learning for the satellite image classification.❍