ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN-TIN

LUẬN VĂN TỐT NGHIỆP ĐẠI HỌC

Development of Yolo machine learning model and real-time streaming operational parameters of CO2 micro algae capture pilot

GIẢNG VIÊN HƯỚNG DẪN:
TS. HOÀNG VĂN HÀ
TS. TRỊNH NGỌC TRUNG

—o0o—

SINH VIÊN THỰC HIỆN:
ĐÀO THANH NGUYÊN-20280068

TP. HỒ CHÍ MINH, 7/2024

Lời cam đoan

...

Lời cảm ơn / Lời ngỏ

Để hoàn thành kì đề cương luận văn này, tôi tỏ lòng biết ơn sâu sắc đến tiến sĩ Trịnh Ngọc Trung và tiến sĩ Hoàng Văn Hà đã hướng dẫn tận tình trong suốt quá trình nghiên cứu.

Tôi chân thành cám ơn quý thầy, cô trong khoa Toán-Tin, Trường đại học khoa học tự nhiên thành phố Hồ Chí Minh đã tận tình truyền đạt kiến thức trong những năm tôi học tập ở trường.

Cuối cùng, tôi xin chúc quý thầy, cô dồi dào sức khỏe và thành công trong sự nghiệp cao quý.

Tóm tắt nội dung

Nội dung chính của luận văn nhằm tìm hiểu, nghiên cứu xây dựng hệ thống sử dụng các mô hình học máy và điện toán đám mây vào theo dõi quá trình nuôi vi tảo thời gian thực trong công nghiệp dựa trên những công trình, công nghệ mới được nghiên cứu và phát triển trong những năm gần đây. Trong quá trình nghiên cứu, chúng tôi đã tiến hành tổng hợp, đánh giá ưu và nhược điểm của cách phương pháp, công nghệ đã và đang được nghiên cứu, sử dụng. Tiếp cận vấn đề theo nhiều hướng khác nhau, chúng tôi thực hiện một số phương pháp sử dụng máy học để dự đoán nồng độ chất, thị giác máy tính để nhận diện bọt khí và xây dựng nền tảng theo dõi thời gian thực trong quá trình nuôi vi tảo. Bên cạnh việc hoàn thành nội dung của đề tài, nhóm chúng tôi đã nghiên cứu thêm một số phần để từ đó đặt nền móng cho các nghiên cứu sau này. Phần còn lại của luận văn tập trung vào việc đánh giá mô hình,xây dựng hệ thống và kết quả đạt được, đồng thời phân tích ưu nhược điểm của mô hình và hệ thống thực hiện và thảo luận những vấn đề mà mô hình và hệ thống còn gặp phải. Cuối cùng, nhóm chúng tôi đề xuất hướng phát triển tiếp theo của đề tài trong tương lai.

Mục lục

Danh sách hình vẽ

Chương 1

Giới thiệu tổng quan vấn đề

1.1 Đặt vấn đề

Trong lĩnh vực công nghệ sinh học, việc áp dụng các tiến bộ khoa học vào quá trình nuôi cấy vi tảo đang mở ra những cơ hội mới cho sản xuất bền vững và phát triển các hóa chất giá trị cao. Đặc biệt, việc kiểm soát quá trình nuôi cấy vi tảo thông qua công nghệ tiên tiến như học máy, thị giác máy tính và nền tảng điện toán đám mây, đang chứng minh là bước đột phá trong việc tối ưu hóa hiệu quả sản xuất và quản lý chất lượng sản phẩm.

Quá trình nuôi vi tảo giúp tạo ra hóa chất AAA, sử dụng thiết bị raman để đo tín hiệu điện kết hợp mô hình học máy, cho phép chúng ta dự đoán nồng độ hóa chất AAA một cách chính xác theo thời gian thực. Điều này không chỉ giúp tối ưu hóa quá trình sản xuất mà còn tăng cường khả năng kiểm soát chất lượng sản phẩm một cách linh hoạt và hiệu quả.

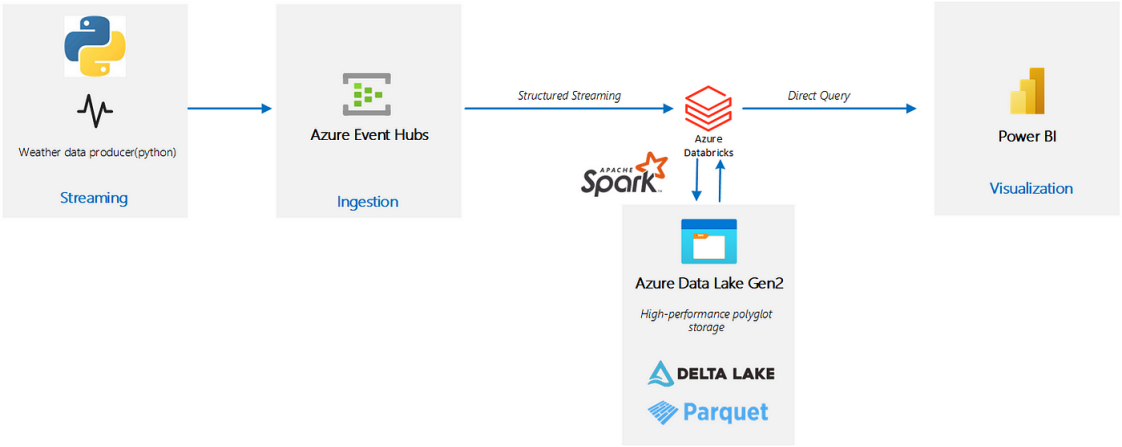
Bằng cách tập chung vào kiểm soát lưu lượng bọt khí CO2 và nồng độ hóa chất AAA trong quá trình nuôi cấy vi tảo, chúng ta có thể đạt được mục tiêu sản xuất một cách nhanh chóng và chính xác mà không cần phải phụ thuộc vào các phương pháp phân tích offline tốn kém và mất thời gian. Sự kết kệt giữa mô hình Yolo và nền tảng Azure cloud mở ra khả năng giải quyết các thách thức trong việc tối ưu hóa quá trình vận hành, đồng thời nâng cao tính linh hoạt và chất lượng trong kiểm soát quá trình sản xuất.

Như vậy, việc áp dụng công nghệ tiên tiến như học máy, thị giác máy tính và nền tảng điện toán đám mấy Azure trong quá trình nuôi cấy vi tảo không chỉ là bước tiến quan trọng trong việc sản xuất hóa chất một cách bền vững mà còn là minh chứng cho sự tiến bộ trong lĩnh vực công nghệ sinh học, mở ra hướng đi mới cho việc tối ưu hóa và kiểm soát quá trình sản xuất.

1.2 Giới thiệu real-time platform trên azure.

Real-time platform trên Azure là một giải pháp công nghệ mạnh mẽ, được thiết kế để xử lý và phân tích dữ liệu trong thời gian thực, mang lại khả năng phản hồi nhanh chóng và hiệu quả cho các ứng dụng và dịch vụ. Sử dụng nền tảng điện toán đám mây Azure, platform này cung cấp một hệ thống linh hoạt và đáng tin cậy cho việc triển khai các giải pháp xử lý dữ liệu thời gian thực.

Cấu trúc



Hình 1.1: Real-Time Data Streaming With Spark Structured Streaming in Azure

Real-time platform trên Azure được xây dựng dựa trên một số thành phần chính sau:

- **Azure Event Hubs:** Đây là dịch vụ xử lý sự kiện quy mô lớn, có khả năng thu thập dữ liệu sự kiện từ hàng triệu thiết bị với độ trễ thấp và băng thông cao. Event Hubs là điểm vào chính cho dữ liệu thời gian thực vào hệ thống.
- **Azure Databricks:** là một nền tảng phân tích dữ liệu dựa trên Apache Spark, nền tảng này cung cấp môi trường giúp xử lý dữ liệu phức tạp, xây dựng và huấn luyện mô hình máy học một cách hiệu quả.
- **Azure Data Lake Gen2:** là một dịch vụ lưu trữ dữ liệu cấp cao. Gen2 kết hợp tính năng của Hadoop Distributed File System (HFDS) với khả năng mở rộng và tính bền vững của Azure Blob Storage, tạo ra một nền tảng lưu trữ dữ liệu linh hoạt và mạnh mẽ.
- **Power BI:** là công cụ trực quan hóa dữ liệu mạnh mẽ của Microsoft, cho phép người dùng kết nối, biến đổi và trực quan hóa dữ liệu từ nhiều nguồn khác nhau một cách dễ dàng. Power BI cung cấp các tính năng trực quan hóa dữ liệu đa dạng như biểu đồ, bảng, bản đồ và dashboard, giúp người dùng hiểu rõ hơn về dữ liệu và đưa ra quyết định thông minh dựa trên thông tin được trực quan hóa một cách rõ ràng.

Chức năng

- **Thu thập dữ liệu thời gian thực:**Từ các thiết bị IoT, ứng dụng web/mobie, và các nguồn dữ liệu khác, đảm bảo dữ liệu được thu thập một cách liên tục và đáng tin cậy.
- **Phân tích dữ liệu thời gian thực:** Phân tích và xử lý dữ liệu ngay lập tức, cho phép phát hiện mẫu, xu hướng và cảnh báo sớm một cách nhanh chóng.
- **Tích hợp và tự động hóa:** Dễ dàng tích hợp với các dịch vụ và ứng dụng khác, tự động hóa các quy trình xử lý dữ liệu và phản hồi.
- **Độ linh hoạt và mở rộng:** Cung cấp khả năng mở rộng tự động, cho phép xử lý lượng dữ liệu lớn mà không cần lo lắng về cơ sở hạ tầng.

Real-time platform trên Azure giúp các tổ chức tận dụng sức mạnh của dữ liệu thời gian thực để tối ưu hóa quy trình vận hành, cải thiện trải nghiệm người dùng và đưa ra quyết định kinh doanh một cách nhanh chóng, kịp thời và chính xác. Điều này không chỉ giúp các doanh nghiệp duy trì lời thế cạnh tranh mà còn tối ưu hóa hiệu suất và giảm thiểu chi phí vận hành.

Nền tảng giúp cung cấp một giải pháp toàn diện cho việc xử lý và phân tích dữ liệu thời gian thực, từ thu thập dữ liệu, xử lý, phân tích, cho đến lưu trữ và trực quan hóa. Điều này giúp các tổ

chức phản ứng nhanh chóng với các sự kiện thời gian thực, tự động hóa quy trình và tạo ra giá trị từ dữ liệu một cách hiệu quả.

Ưu điểm

- **Xử lý dữ liệu thời gian thực:** Platform này cho phép xử lý và phân tích dữ liệu trong thời gian thực, giúp tổ chức phản ứng nhanh chóng với các sự kiện và tình huống cần giải quyết ngay lập tức.
- **Phản hồi nhanh chóng:**Real-time platform trên Azure mang lại khả năng phản hồi nhanh chóng và hiệu quả, giúp cải thiện trải nghiệm người dùng và quyết định kinh doanh.
- **Tích hợp linh hoạt:**Platform này dễ dàng tích hợp với các dịch vụ và ứng dụng khác trên nền tảng Azure, tạo ra một hệ sinh thái phân tích dữ liệu toàn diện.
- **Mở rộng dễ dàng:** Real-time platform trên Azure cung cấp khả năng mở rộng tự động, cho phép xử lý lượng dữ liệu lớn mà không cần lo lắng về cơ sở hạ tầng.

Nhược điểm

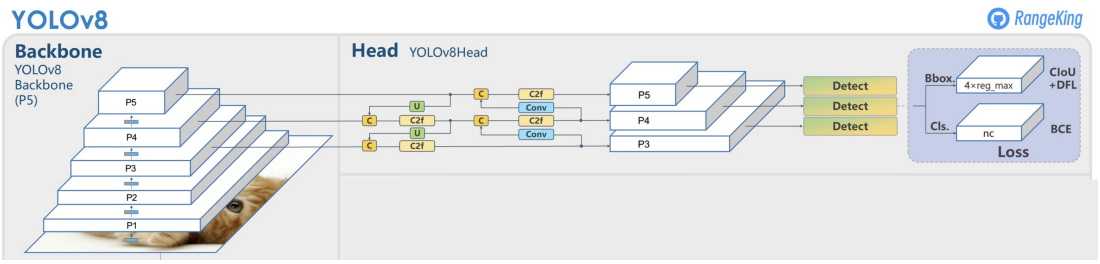
- **Chi phí:**Sử dụng các dịch vụ xử lý dữ liệu thời gian thực có thể tạo ra chi phí cao, đặc biệt khi xử lý lượng dữ liệu lớn và cần sự phản hồi nhanh.
- **Độ phức tạp:**Cấu hình và quản lý real-time platform trên Azure có thể đòi hỏi kiến thức chuyên sâu về phân tích dữ liệu và các công nghệ liên quan, đôi khi tạo ra thách thức cho người mới sử dụng.
- **Yêu cầu chuyên môn:** Để tận dụng hết tiềm năng của platform, người dùng cần có kiến thức vững về xử lý dữ liệu thời gian thực và phân tích dữ liệu, có thể tạo ra rào cản cho người mới bắt đầu.

1.3 Các phương pháp machine learning và statistic sử dụng.

1.3.1 Partial Least Squares Regression (PLSR)

PLS-Regression là phương pháp PLS đơn giản nhất, trong hóa học và công nghệ được sử dụng nhiều nhất ở dạng hai khối PLS dự đoán. PLSR là một phương pháp liên kết hai ma trận dữ liệu X và Y, theo mô hình đa biến tuyến tính, nhưng vượt xa hồi quy truyền thống ở chỗ mô hình hóa cấu trúc của X và Y. PLSR có ưu điểm từ khả năng phân tích dữ liệu có nhiều biến nhiễu, cộng tuyến và thậm chí không đầy đủ trong cả X và Y. PLSR mong đợi rằng độ chính xác của các tham số mô hình được cải thiện khi số lượng biến và quan sát liên quan ngày càng tăng. PLSR đã phát triển trở thành một công cụ tiêu chuẩn trong hóa học và được sử dụng trong hóa học và kỹ thuật.[1]

1.3.2 Tổng quan về Yolov8



Hình 1.2: YOLOv8 Architecture

YOLOv8 là phiên bản mới của mô hình phát hiện đối tượng YOLO. Phiên bản này có kiến trúc giống như các phiên bản tiền nhiệm nhưng có nhiều cải tiến so với các phiên bản trước của YOLO, chẳng hạn như kiến trúc mạng thần kinh mới sử dụng cả Feature Pyra-mid Network (FPN) và Path Aggregation Network(PAN) và một công cụ ghi nhãn mới giúp đơn giản hóa quá trình chú thích. Công cụ ghi nhãn này chứa một số tính năng hữu ích như ghi nhãn tự động, phím tắt ghi nhãn và phím nóng có thể tùy chỉnh. Sự kết hợp của các tính năng này giúp việc chú thích hình ảnh để huấn luyện mô hình trở nên dễ dàng hơn.[2]

FPN hoạt động bằng cách giảm dần độ phân giải không gian của hình ảnh đầu vào đồng thời tăng số lượng kênh đặc trưng. Điều này dẫn đến việc tạo ra các bản đồ đặc trưng có khả năng phát hiện các vật thể ở các tỷ lệ và độ phân giải khác nhau. Mặt khác, kiến trúc PAN tổng hợp các tính năng từ các cấp độ khác nhau của mạng thông qua việc bỏ qua các kết nối. Bằng cách đó, mạng có thể nắm bắt tốt hơn các đặc điểm ở nhiều tỷ lệ và độ phân giải, điều này rất quan trọng để phát hiện chính xác các vật thể có kích thước và hình dạng khác nhau.

1.4 Mục tiêu của đề tài

Mục tiêu của đề tài là nghiên cứu, hiểu và hiện thực một số phương pháp hiệu quả để xây dựng một hệ thống theo dõi thời gian thực hiệu quả.

Một số vấn đề đặt ra:

- Làm thế nào để giải quyết bài toán trên?
- Cách tiếp cận như thế nào?
- Những công nghệ nào đã và hiện đang được sử dụng?
- Tối ưu hóa chi phí vận hành như thế nào?
- Hướng cải tiến?

Như vậy để thực hiện theo đúng mục tiêu của đề tài cần xác định một số công việc phải giải quyết như sau:

- Tìm kiếm và thu thập dữ liệu phù hợp với nội dung đề tài.
- Tìm hiểu các phương pháp tiếp cận đã được hiện thực
- Lựa chọn mô hình phù hợp
- Lựa chọn kiến thức hệ thống phù hợp
- Lên kế hoạch hiện thực, phát triển hệ thống theo dõi thời gian thực.

1.5 Cấu trúc luận văn

Nội dung của luận văn sẽ được trình bày trong những chương sau:

- Chương 1: Giới thiệu tổng quan vấn đề
- Chương 2: Xây dựng real-time platform
- Chương 3: Mô hình Yolo và ứng dụng
- Chương 4: Kết quả thí nghiệm
- Chương 5: Tổng kết, đánh giá và định hướng kế hoạch phát triển.

Chương 2

Xây dựng real-time platfrom trên Azure

2.1 Xây dựng mô hình dự đoán dựa trên tín hiệu điện

2.1.1 Chuẩn bị dữ liệu

2.1.2 Xử lý dữ liệu

2.1.3 Thuật toán Partial Least Squares Regression (PLSR)

2.1.4 Tiến hành huấn luyện và đánh giá mô hình

2.2 Tiến hành xây dựng real-time platform

2.2.1 Thành phần chính

2.2.2 Kiến trúc và nguyên lý hoạt động

2.2.3 Đánh giá chi phí vận hành

2.3 Thử nghiệm và kết luận

Chương 3

Mô hình Yolo và Ứng dụng

3.1 Kiến thức nền tảng

3.2 Mô hình Yolo

3.3 Thuật toán chính

3.4 Xây dựng ứng dụng phát hiện bọt khí trong nuôi vi tảo bằng Yolo

Chương 4

Thực nghiệm và đánh giá

4.1 ...

4.2 ...

4.3 ...

Chương 5

Tổng kết, đánh giá và định hướng kế hoạch phát triển.

5.1 ...

5.2 ...

5.3 ...

Tài liệu tham khảo

[1] Svante Wold, Michael Sjostrom, Lennart Eriksso, PLS-regression: a basic tool of chemometrics, 2001

[2] Dillon Reis, Jordan Kupec, Jacqueline Hong, Ahmad Daoudi Georgia Institute of Technology, Real-Time Flying Object Detection with YOLOv8,2023
https://www.semanticscholar.org/reader/231a434f8fac0b01cbc05890b283f4d9da4cb100