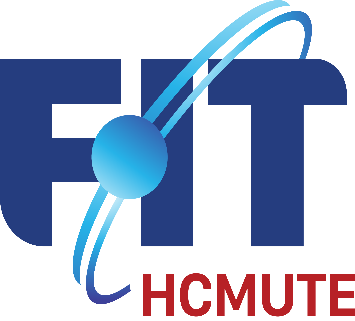
Bộ GiÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH

KHOA CÔNG NGHỆ THÔNG TIN

BỘ MÔN ĐIỆN TOÁN ĐÁM MÂY

🞠◊🞠◊🞠



ĐOÀN THANH NGUYÊN: 19110407

TRẦN NHẤT QUANG: 19110242

LẠI QUANG PHÁT: 19110424

Đề tài:

TÌM HIỂU VỀ AMAZON ELASTIC INFERENCE

GIÁO VIÊN HƯỚNG DẪN

TS. Huỳnh Xuân Phụng

KHOÁ 2019 – 2023

Mục lục

[1. Lý thuyết về Amazon Elastic Inference 3](#_Toc117240990)

[**1.1 Giới thiệu** 3](#_Toc117240991)

[**1.2 Tích hợp với Amazon SageMaker, Amazon EC2 và Amazon ECS** 3](#_Toc117240992)

[**1.3 Hỗ trợ TensorFlow, Apache MXNet và PyTorch** 3](#_Toc117240993)

[**1.4 Mở hỗ trợ định dạng Neural Network Exchange (ONNX)** 4](#_Toc117240994)

[**1.5 Lựa chọn các hoạt động chính xác đơn lẻ hoặc hỗn hợp** 4](#_Toc117240995)

[**1.6 Có sẵn với nhiều mức tăng tốc** 4](#_Toc117240996)

[**1.7 Tự động điều chỉnh tỷ lệ** 5](#_Toc117240997)

# **Lý thuyết về Amazon Elastic Inference**

## **1.1 Giới thiệu**

Amazon Elastic Inference cho phép bạn đính kèm lượng tăng tốc suy luận do GPU hỗ trợ vừa phải vào bất kỳ phiên bản Amazon EC2, phiên bản Amazon SageMaker hoặc tác vụ ECS nào. Điều này có nghĩa là bây giờ bạn có thể chọn phiên bản CPU phù hợp nhất với nhu cầu tính toán, bộ nhớ và lưu trữ tổng thể của ứng dụng, sau đó định cấu hình riêng lượng tăng tốc suy luận do GPU hỗ trợ mà bạn cần.

Lợi ích :

Giảm chi phí inference lên tới 75%

Hỗ trợ các mô hình TensorFlow và Apache MXNet, với các framework bổ sung sắp ra mắt.

Có thể cung cấp ít nhất là một TFLOPS chính xác đơn (nghìn tỷ phép toán dấu phẩy động mỗi giây) tăng tốc suy luận hoặc nhiều nhất là 32 TFLOPS chính xác hỗn hợp.

Dễ dàng mở rộng phạm vi tăng tốc inference lên và xuống bằng cách sử dụng các nhóm Auto Scaling của Amazon EC2 để đáp ứng nhu cầu của ứng dụng của bạn mà không cần cung cấp quá mức dung lượng.

## **1.2 Tích hợp với Amazon SageMaker, Amazon EC2 và Amazon ECS**

Có nhiều cách để chạy khối lượng inference workloads trên AWS: triển khai mô hình của bạn trên Amazon SageMaker để có trải nghiệm được quản lý đầy đủ hoặc chạy mô hình đó trên các phiên bản Amazon EC2 hoặc các tác vụ Amazon ECS và tự quản lý. Amazon Elastic Inference được tích hợp để hoạt động liền mạch với Amazon SageMaker, Amazon EC2 và Amazon ECS, cho phép bạn thêm khả năng tăng tốc suy luận trong mọi tình huống. Bạn có thể chỉ định mức tăng tốc suy luận mong muốn khi tạo điểm cuối HTTPS của mô hình trong Amazon SageMaker, khi bạn khởi chạy phiên bản Amazon EC2 và khi bạn xác định tác vụ Amazon ECS của mình.

## **1.3 Hỗ trợ TensorFlow, Apache MXNet và PyTorch**

Amazon Elastic Inference được thiết kế để sử dụng với các phiên bản nâng cao của AWS của TensorFlow Serving, Apache MXNet và PyTorch. Những cải tiến này cho phép các framework công tác tự động phát hiện sự hiện diện của bộ tăng tốc suy luận, phân phối tối ưu các hoạt động mô hình giữa GPU của bộ tăng tốc và CPU của phiên bản và kiểm soát an toàn quyền truy cập vào bộ tăng tốc của bạn bằng chính sách AWS Identity and Access Management (IAM). Các thư viện TensorFlow Serving, MXNet và PyTorch nâng cao được cung cấp tự động trong Amazon SageMaker, AWS Deep Learning AMIs và AWS Deep Learning Containers, vì vậy bạn không phải thực hiện bất kỳ thay đổi mã nào để triển khai các mô hình của mình trong sản xuất.

## **1.4 Mở hỗ trợ định dạng Neural Network Exchange (ONNX)**

ONNX là một định dạng mở giúp bạn có thể đào tạo một mô hình trong một khuôn khổ học tập sâu và sau đó chuyển nó sang một khuôn khổ khác để suy luận. Điều này cho phép bạn tận dụng các điểm mạnh tương đối của các khuôn khổ khác nhau. ONNX được tích hợp vào PyTorch, MXNet, Chainer, Caffe2 và Bộ công cụ nhận thức của Microsoft, và có các trình kết nối cho nhiều khung công tác khác bao gồm TensorFlow. Để sử dụng các mô hình ONNX với Amazon Elastic Inference, các mô hình được đào tạo của bạn cần được chuyển sang phiên bản Apache MXNet được AWS tối ưu hóa để triển khai sản xuất.

## **1.5 Lựa chọn các hoạt động chính xác đơn lẻ hoặc hỗn hợp**

Máy gia tốc Amazon Elastic Inference hỗ trợ cả hoạt động chính xác đơn (dấu chấm động 32 bit) và hoạt động chính xác hỗn hợp (dấu chấm động 16 bit). Độ chính xác đơn cung cấp một phạm vi số cực kỳ lớn để đại diện cho các thông số được sử dụng bởi mô hình của bạn. Tuy nhiên, hầu hết các mô hình không thực sự cần những con số chính xác và tính toán lớn như vậy dẫn đến việc giảm hiệu suất không cần thiết. Để tránh vấn đề đó, các phép toán có độ chính xác hỗn hợp cho phép bạn giảm một nửa phạm vi số để đạt được hiệu suất suy luận cao hơn tới 8x.

## **1.6 Có sẵn với nhiều mức tăng tốc**

Amazon Elastic Inference có sẵn ở nhiều kích thước thông lượng khác nhau, từ 1 đến 32 nghìn tỷ hoạt động dấu phẩy động mỗi giây (TFLOPS) trên mỗi bộ gia tốc, giúp tăng tốc hiệu quả cho một loạt các mô hình suy luận bao gồm thị giác máy tính, xử lý ngôn ngữ tự nhiên và nhận dạng giọng nói. So với phiên bản Amazon EC2 P3 độc lập bắt đầu ở 125 TFLOPS (phiên bản P3 nhỏ nhất hiện có), Amazon Elastic Inference bắt đầu ở một TFLOPS duy nhất trên mỗi bộ gia tốc. Điều này cho phép bạn mở rộng quy mô gia tốc suy luận theo từng bước phù hợp hơn. Bạn cũng có thể chọn từ các kích thước bộ gia tốc lớn hơn, lên đến 32 TFLOPS cho mỗi bộ gia tốc, cho các mô hình phức tạp hơn.

## **1.7 Tự động điều chỉnh tỷ lệ**

Amazon Elastic Inference có thể là một phần của cùng một nhóm Tự động mở rộng quy mô Amazon EC2 mà bạn sử dụng để mở rộng quy mô các phiên bản Amazon SageMaker, Amazon EC2 và Amazon ECS của mình. Khi EC2 Auto Scaling thêm nhiều phiên bản EC2 hơn để đáp ứng nhu cầu của ứng dụng của bạn, nó cũng mở rộng quy mô bộ gia tốc được gắn vào mỗi phiên bản. Tương tự, khi Tự động mở rộng quy mô làm giảm các phiên bản EC2 của bạn khi nhu cầu đi xuống, nó cũng giảm tỷ lệ bộ gia tốc kèm theo cho mỗi trường hợp. Điều này giúp bạn dễ dàng mở rộng khả năng tăng tốc suy luận cùng với khả năng tính toán của ứng dụng để đáp ứng nhu cầu của ứng dụng của bạn.

1. **Đồ Án**

Tên đề tài : Viết ứng dụng sử dụng dịch vụ Amazon Elastic Inference ( cụ thể là ứng dụng Machine learning phân loại ảnh trên tập dữ liệu Caltech-256)

Link github: *https://github.com/TranNhatQuang2405/Amazon\_Elastic\_Inference\_Project*

Service có thể sử dụng: ***low-cost GPU-powered acceleration to Amazon EC2 and Amazon SageMaker instances***, ***hỗ trợ TensorFlow, Apache MXNet và PyTorch***