

MLE Class

Scale Machine Learning System

Yêu cầu đầu vào:

- Học viên [lớp học TF](#) (hoặc kiến thức tương đương) + [lớp học Python \(miễn phí\)](#).

Hình thức học:

- Video (Lưu trữ lý thuyết). Đăng tải 19:00 **thứ 2**.
- Zoom (Trao đổi, chữa bài tập, giải đáp thắc mắc, kiến thức nâng cao)
 - Thời gian: 19:00 - 21:00 **Thứ năm** hàng tuần
- Video/Slide/Bài tập lưu trữ trên hệ thống:
<https://protonx.io/courses/6242c8a0bc6a6b00190d3903>
- Blog tham khảo: <https://scale-ml.github.io/>

Đăng ký MLEs 02: <https://protonx.io/courses/635a68c5fd9acb001a858aca>

Bài toán của học viên:

- Trích xuất từ khóa từ văn bản (Keyword extraction)
- Tối ưu hóa tốc độ LSTM cho dữ liệu Timeseries.
- Tối ưu hóa tốc độ cho mô hình phân đoạn ảnh (Image Segmentation)
- Mô hình Segmentation trong nhà máy - hệ thống vận hành này áp dụng cho tầm 10-50 máy vision. Mỗi máy vision có từ 2-4 camera, mỗi camera đưa ảnh về có thể xài 1-2 model DL, Cam có Cam không xài DL.
- Phát hiện anomaly trong dữ liệu time series gửi về từ sensor, đưa ra quyết định real time dựa vào ML.

3 hướng tiếp cận xây dựng hệ thống học máy lớn

- **Nguyên tắc 1:** Sử dụng kỹ thuật lấy mẫu con (**Sub Sampling**) hiệu quả
- **Nguyên tắc 2:** Sử dụng các mô hình/kỹ thuật tiên tiến
- **Nguyên tắc 3:** Sử dụng tối đa năng lực tính toán của phần cứng

Buổi	Thời gian	Nội dung	Hình thức (Dự kiến)
	19:00 - 19:45 Thứ hai Ngày 22/08/2022	Gặp gỡ đội ngũ giảng dạy Giới thiệu về lớp học, giảng viên, trao đổi với học viên về ứng dụng AI trong công việc Đường dẫn	Zoom
1	19:00 Thứ hai Ngày 22/08/2022	Ôn tập Học máy. Các thuật toán training. Ôn tập học máy (Video Quizz) Đường dẫn	Video + Quizz + Bài tập lập trình thi đấu

		Hồi quy tuyến tính + Gradient Descent (Video Quizz Lập trình thi đấu) Đường dẫn	
	19:00 - 21:00 Thứ năm Ngày 25/08/2022	Ôn tập các mô hình cơ bản - Logistic Regression - Softmax Regression - Mạng nơ ron - CNN - RNN - Giới thiệu qua Transformer - Ứng dụng Transformer Video	Zoom
2	19:00 Thứ hai Ngày 29/08/2022	Ôn tập xác suất • Giới thiệu Đếm • Giới thiệu xác suất • Biến ngẫu nhiên + Phân phối liên tục • Phân phối chuẩn	Video + Quizz + Bài tập lập trình thi đấu
3	19:00 - 21:00 Thứ năm Ngày 01/09/2022	Ôn tập lập trình mô hình - Lập trình bài toán nhận diện hình ảnh + dự đoán âm thanh + sinh từ với Tensorflow - Họ Transformer và các mô hình ứng dụng: Bert/GPT Giới thiệu Pytorch - Các hàm cơ bản - Xây dựng mô hình cơ bản với Pytorch - Dataset và Dataloader trong Pytorch - Custom Training với Pytorch Thư viện Hugging Face - Tokenizer trong HuggingFace - Ý tưởng BPE (Byte Pair Encoding) - Thư viện Hugging Face cho NLP - Bài toán phân loại trên Tiếng Việt Notebook Pytorch Phân loại văn bản tiếng Việt với HuggingFace	Zoom

4		Limit Theorems - Concentration inequalities - Bất đẳng thức Markov. Video - Bất đẳng thức Chebyshev Video - Luật số lớn yếu (Weak Law of Large Numbers) - Luật số lớn mạnh (Strong Law of Large Numbers) Video - Áp dụng lấy mẫu để tính toán tối ưu cho chi phí rủi ro thực nghiệm (Empirical Risk Minimization) Video - Định luật giới hạn trung tâm (Central Limit Theorem - CLT) Slide Video Notebook	Video + Quizz + Bài tập lập trình thi đấu
6		[Live Coding] Làm việc và lưu trữ, theo dõi dữ liệu với MongoDB. Ứng dụng các định lý vào bài toán thực tiễn - Tối ưu quá trình training sử dụng sub sampling Hướng dẫn setup MongoDB - Đăng ký tài khoản MongoDB Atlas - Các cách kết nối MongoDB. Doc - Thư viện PyMongo - Liệt kê, thêm sửa, xoá dữ liệu Thiết kế Pipeline clean dữ liệu thô và theo dõi sự thay đổi của dữ liệu Làm việc và kiểm tra chất lượng dữ liệu - TDD (Test-Driven Development) và quy trình review test/code. - Kiểm thử quá trình tiền xử lý dữ liệu - Thư viện unittest. Video Project Code Sách MongoDB cài đặt MiniConda	Zoom
7		Các kỹ thuật training hiệu quả 1 (Nguyên tắc 2) - Tối ưu hàm lồi (Convex Optimization) - Sử dụng các thuật toán training. AdaGrad/Adam - Cài đặt các tham số khởi tạo Slide Video Notebook	Video + Quizz + Bài tập lập trình thi đấu

8		[Live Coding] Phân bổ dữ liệu để training. Đặt giờ tự động các tác vụ khác nhau - Giới thiệu Google Cloud Storage - Luyện tập sử dụng các thuật toán khác nhau để tối ưu hiệu quả training. - Đặt giờ tự động pipeline - Kéo dữ liệu - Clean dữ liệu - Tự động training. Slide Video Code	Zoom
9		Nghỉ giữa khóa - Team giảng dạy review lại nội dung - Khảo sát đóng góp của học viên nhằm nâng cao chất lượng	
10		Giảm chiều (Nguyên tắc 2) - Tại sao phải giảm chiều - Bậc của ma trận (Matrix Rank) - Thuật toán SVD (Singular Value Decomposition) Slide Video Notebook	Video + Quizz + Bài tập lập trình thi đấu
11		Các kỹ thuật training hiệu quả 2 (Nguyên tắc 2) Nâng cao: PCA - Principle Component Analysis Slide Video Notebook Bài tập	Zoom
12		Một số kỹ thuật khác Grid Search và Random Search - Điểm mạnh + yếu từng phương pháp - Giới thiệu Keras Tuner Slide Video	
13		CPUs - Giới thiệu CPU - Tính toán song song trên CPUs - Nhiều lệnh song song (ILP) - Một lệnh trên nhiều luồng dữ liệu (SIMD) - Đa luồng (Multi-thread parallelism) - Phân tán (Distributed Compute) - Đa luồng tiên xử lý dữ liệu - Bài tập: Lập trình multi threading. Slide Video Notebook	Zoom

14		Đồng bộ và bất đồng bộ - Phân biệt hai khái niệm này - SGD Bất đồng bộ / Thuật toán HOGWILD HOGWILD!: A Lock-Free Approach to Parallelizing Stochastic Gradient Descent: https://proceedings.neurips.cc/paper/2011/file/218a0aefd1d1a4be65601cc6ddc1520e-Paper.pdf https://people.ece.ubc.ca/matei/papers/ Slide Video Notebook	
15		Chữa bài tập + hướng dẫn thuê/sử dụng VPS CPUs trên Digital Ocean + Google Cloud - Bài tập Pytorch - Bài tập HuggingFace - Bài tập MongoDB - Hướng dẫn thuê/sử dụng CPUs trên Digital Ocean + Google Cloud - Set up pipeline trên VPS Slide Video Notebook Quizz 1 Quizz 2 Quizz 3 Quizz 4 demo setup VPC.	
16		Cấu trúc Ram - L1/L2/L3 Cache - Không gian bộ nhớ vs băng thông bộ nhớ (Memory Locality vs Memory Bandwidth) Video	Zoom
17		Deploy CPUs server - Kỹ thuật Prefetching trong RAM - Giới thiệu về Flask, Webserver và deploy một model trên CPU Đọc thêm: - Prefetching https://www.tensorflow.org/guide/data_performance#prefetching Slide Quizz Code Video	Zoom
18		Machine learning CI/CD - Jenkins - Set up CI/CD sử dụng Travis CI - Auto Deployment Slide Video Quizz Code	Zoom

19		Sử dụng Docker - Sử dụng Docker: Docker Image + Docker Container - Giới thiệu Docker Compose Slide Video Quizz	Video
20		Google Cloud + Docker cần thiết cho học máy - Cloud Storage: Lưu trữ data lớn + Model lớn - Load Balancing: Phân luồng Máy chủ - Compute Engine: Máy chủ - Instance Group: Tự động scaling máy chủ Slide Video Code	Zoom
21		MLOps và những điều cần biết - Đọc bài và trả lời các câu hỏi liên quan tới MLOps - Khái niệm MLOps (ML operations) - Tổng quan về End-to-End Machine Learning Workflow - Nguyên tắc MLOps	
22		GPUs và TPUs - Sự khác biệt giữa CPUs và GPUs (tối ưu cho độ trễ và tối ưu cho băng thông) - Giới thiệu CUDA - Graphic Pipelines - Demo training trên A100 - Song song trên GPU So sánh GPUs và TPUs - Điểm mạnh của TPUs so với GPUs - TPU paper: https://arxiv.org/pdf/1704.04760.pdf - Quizz 1 - Quizz 2 - Quizz 3 Giới thiệu thư viện Numba + Cupy - Lập trình song song trên GPUs Slide Video Code	Zoom (Thứ 5 ngày 10/11/2022)
23		Floating Point Format - Giới thiệu IEEE-754 Floating-Point Standard - TFLOPs Mix-precision: - https://www.tensorflow.org/guide/mixed_precision	Zoom

		Đọc nghiên cứu - Deep Learning with Limited Numerical Precision: http://proceedings.mlr.press/v37/gupta15.pdf - Map-Reduce for Machine Learning on Multicore: https://proceedings.neurips.cc/paper/2006/file/77ee3bc58ce560b86c2b59363281e914-Paper.pdf Học máy chính xác thấp (Low-Precision Machine Learning) - Low-Precision Machine Learning - Brain floating point - Bfloat16 - Mixed Precision Training Các dạng lượng tử hóa hậu training: - Quantization hậu training với float16 - Quantization hậu training với khoảng động - Quantization hậu training với số nguyên Demo training mô hình với FP16 cùng Tensorflow. https://www.tensorflow.org/guide/mixed_precision Đọc nghiên cứu: Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference https://arxiv.org/pdf/1712.05877.pdf Video Slide Quizz	
24		Xây dựng Crawler/Spider Web - Chi tiết về HTML/CSS - Xử lý text từ website - Build Text Dataset Quizz Đăng sau Dataset - Scan Orders (Sắp xếp dữ liệu tối ưu khi train mô hình học máy trên Ram) - Đánh giá hiệu năng của Tensorflow TF Data Pipeline.	

25		Kỹ thuật tăng cường hiệu năng mô hình - Lượng tử hóa mô hình - Lượng tử hóa hậu training - Lượng tử hóa có thể training - Tỉa mô hình (Pruning) - Lượng tử hóa có thể training - Nén mô hình (Model compression) - Chắt lọc tri thức (Knowledge distillation) - Phân cụm tham số (Weight clustering) Đọc nghiên cứu - DEEP COMPRESSION: COMPRESSING DEEP NEURAL NETWORKS WITH PRUNING, TRAINED QUANTIZATION AND HUFFMAN CODING https://arxiv.org/pdf/1510.00149.pdf - WHAT IS THE STATE OF NEURAL NETWORK PRUNING? https://proceedings.mlsys.org/paper/2020/file/d2ddea18f00665ce8623e36bd4e3c7c5-Paper.pdf Video Quizz	Video + Quizz + Bài tập lập trình thi đấu
25		Đào tạo phân tán (Nguyên tắc 3) - Các cơ chế Push/Pull/BroadCast/Reduce/All-reduce/Wait - Parameter server model - Giới thiệu + thực hành tf.distribute.Strategy - Thực hành training mô hình trên nhiều GPUs Dữ liệu song song (Data Parallelism) - Data Parallelism - Distributed Data Parallelism - ZeRO Data Parallelism Video Slide	Zoom
		Tổng kết	Zoom
		Đọc nghiên cứu - - Large Scale Distributed Deep Networks: https://proceedings.neurips.cc/paper/2012/file/6aca97005c68f1206823815f66102863-Paper.pdf - Optimization Methods for Large-Scale Machine Learning: https://arxiv.org/pdf/1606.04838v1.pdf - Scale Deep Learning https://berkeley-deep-learning.github.io/cs294-131-s17/slides/Catanzaro_Berkeley_CS294.pdf	Zoom

		- Tensorflow WhitePaper http://download.tensorflow.org/paper/whitewaterpaper2015.pdf - https://github.com/antonpaquin/SystolicArrayDemo/blob/master/systolic.py -	
--	--	---	--

More References

TPU: <https://cloud.google.com/tpu/docs/system-architecture-tpu-vm>

Zero-Shot Learning in Modern NLP:

<https://joeddav.github.io/blog/2020/05/29/ZSL.html>

Building an Efficient Machine Learning API

<https://www.carted.com/blog/building-an-efficient-machine-learning-api/>