

AI4Code: Understand Code in Python Notebooks

Atlantis \w a bunch of ensemble

Huỳnh Việt Thám¹ Lê Công Luận¹ Huỳnh Hoàng Huy¹ Đoàn Đình Toàn¹ Kiều Vũ Minh Đức¹

¹Department of Computer Science,
Faculty of Information Technology University Of Science

I90 À, có tên nhóm gì hay hay khum a Bùi

20:50

I90 Đang cần một tên nhóm cho nhóm deeplearning :<

20:51

I90 À, có tên nhóm gì hay hay khum a Bùi

pher Sao giờ tui thành tên-nhóm recommendation system rồi 😢

pher hmm

pher Atlantis?

pher Theo nghĩa "nếu bạn học quá sâu bạn sẽ gặp... những thứ bạn không nên gặp?"

fit@hcmus

Các phần chính

① Giới thiệu lại bài toán

fit@hcmus

Các phần chính

- ① Giới thiệu lại bài toán
- ② Một số hướng tiếp cận

fit@hcmus

Các phần chính

- ① Giới thiệu lại bài toán
- ② Một số hướng tiếp cận
- ③ Một số kiến thức tiên nghiệm

fit@hcmus

Các phần chính

- ① Giới thiệu lại bài toán
- ② Một số hướng tiếp cận
- ③ Một số kiến thức tiên nghiệm
- ④ Phương pháp của nhóm

Các phần chính

- ① Giới thiệu lại bài toán
- ② Một số hướng tiếp cận
- ③ Một số kiến thức tiên nghiệm
- ④ Phương pháp của nhóm
- ⑤ Hướng phát triển trong tương lai

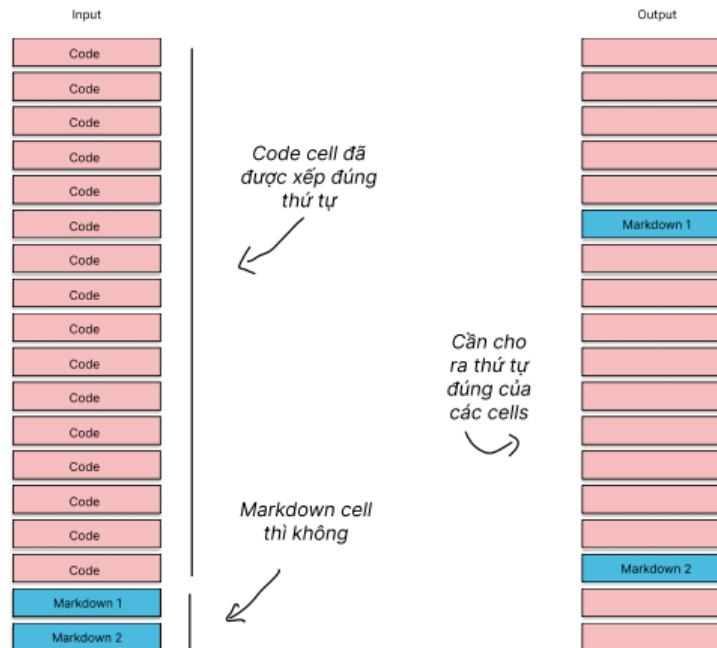
Các phần chính

- ① Giới thiệu lại bài toán
- ② Một số hướng tiếp cận
- ③ Một số kiến thức tiên nghiệm
- ④ Phương pháp của nhóm
- ⑤ Hướng phát triển trong tương lai
- ⑥ Kết luận

Giới thiệu lại bài toán

Giới thiệu lại bài toán

AI4Code: Understand Code in Python Notebooks



fit@hcmus

Figure 1: Minh họa bài toán phiên bản xúc tích

Giới thiệu lại bài toán

AI4Code: Understand Code in Python Notebooks

Bài toán

Cho một nhóm các code cell đã được sắp xếp đúng thứ tự, và một nhóm các markdown cell, tất cả đều cùng nằm trong một notebook. Mô hình cần chèn các markdown cell vào trong các code cell đã được sắp xếp sao cho thứ tự của toàn bộ các cell trả ra đúng với thứ tự của notebook gốc.

fit@hcmus

Giới thiệu lại bài toán

AI4Code: Understand Code in Python Notebooks

Độ đo đánh giá: Kendall tau correlation

Gọi S là số lượng swap các cặp cell kề nhau để chuyển một notebook dự đoán sang một notebook tương ứng trong tập ground-truth. Độ đo Kendall Tau phản ánh cho số lượng swap cần thực hiện được tính như sau:

$$K = 1 - 4 \frac{\sum_i S_i}{\sum_i n_i(n_i - 1)} \quad (1)$$

Với n_i là số lượng cell trên test notebook i .

fit@hcmus

Một số yếu tố khác

- Submission notebook không được phép kết nối với internet, lý do là vì việc kết nối với internet trong inference time có thể khiến test data bị leak hoặc mô hình có thể search các notebook trên database khác để leak dự đoán. Luật này ảnh hưởng tới việc triển khai một số hướng tiếp cận preprocessing như dịch các markdown cell về tiếng Anh.
- Inference time tối đa là 9hrs.

Các hướng tiếp cận

Các hướng tiếp cận

fit@hcmus

Hướng tiếp cận: mô hình pair-wise

Dựa trên solution của yuanzhe zhou

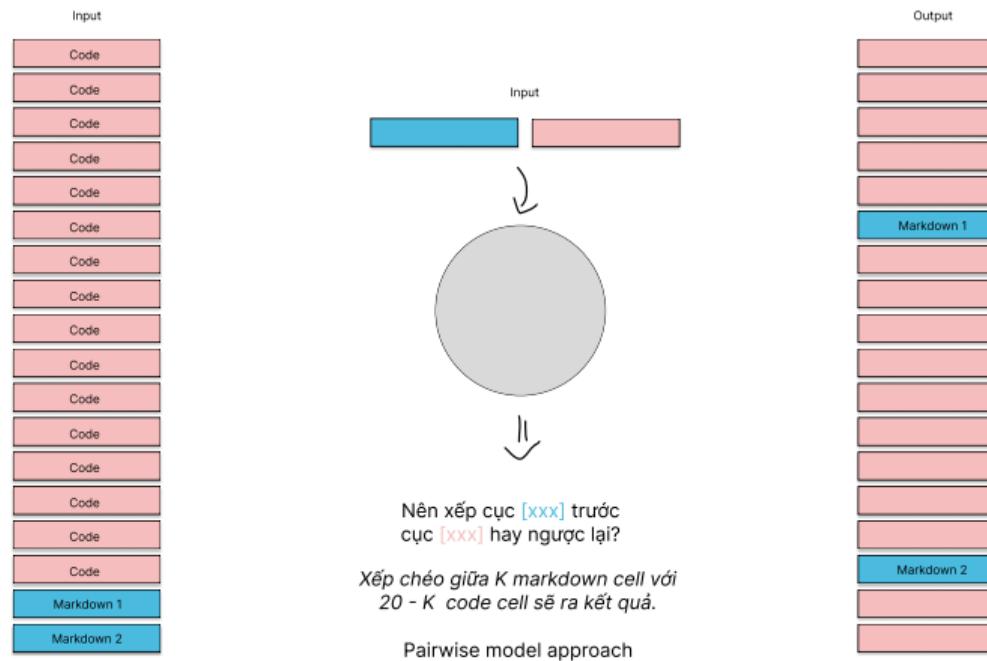


Figure 2: Hướng tiếp cận pair-wise cho ai4code

Hướng tiếp cận: mô hình pair-wise

Dựa trên solution của yuanzhe zhou

Nhận xét

- Hướng tiếp cận pairwise có thể cho phép nhúng thêm các đặc trưng để force prediction (ví dụ như a. sẽ có xu hướng đứng trước b.). Tuy nhiên, việc chọn ra các pair để tiến hành bỏ vào trong pair-wise model sẽ chậm dẫn tới inference time rất nhạy cảm.
- Tuy mô hình nhỏ nhưng việc reproduce lại kết quả lại gặp nhiều khó khăn. Với cùng số epochs nhưng không có các siêu tham số thì việc fine-tune để đạt tới kết quả 0.84x như các public notebook là rất mất thời gian.

fit@hcmus

Hướng tiếp cận: Mô hình point-wise

Dựa trên solution của suicaokhoailang

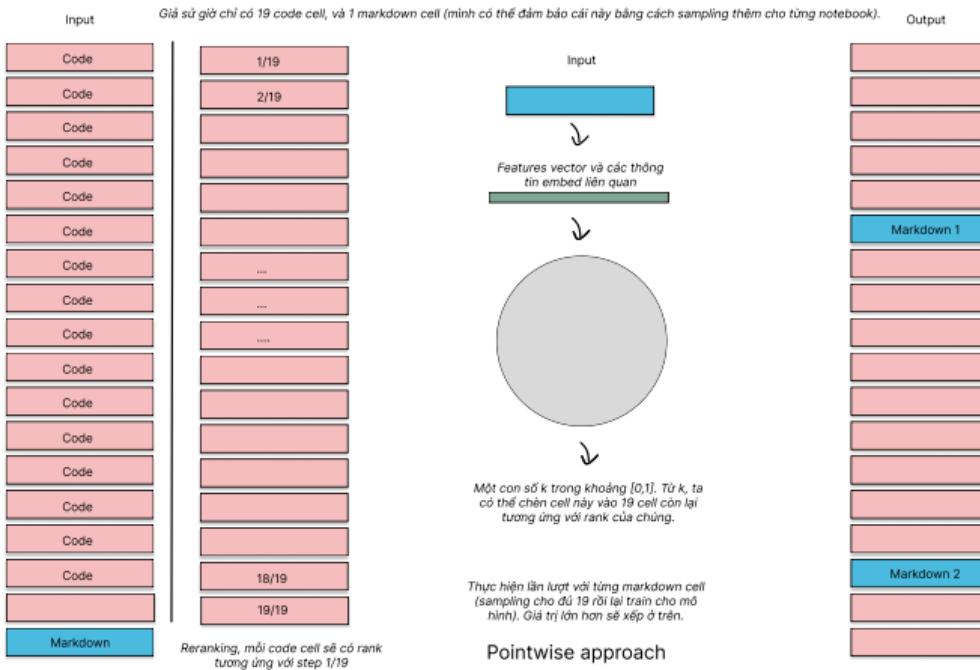


Figure 3: Hướng tiếp cận point-wise cho ai4code

Hướng tiếp cận: Mô hình point-wise

Dựa trên solution của suicaokhoailang

Nhận xét

- Cách tiếp cận này khả thi, thời gian inference thấp vì chỉ predict trên các markdown cell.
- Thời gian fine-tune lâu.

fit@hcmus

Những gì nhóm có

Phần cứng nhóm:

- 1x GTX 3080 (laptop)
- 2x GTX 2080 (server)

fit@hcmus

Những gì nhóm có

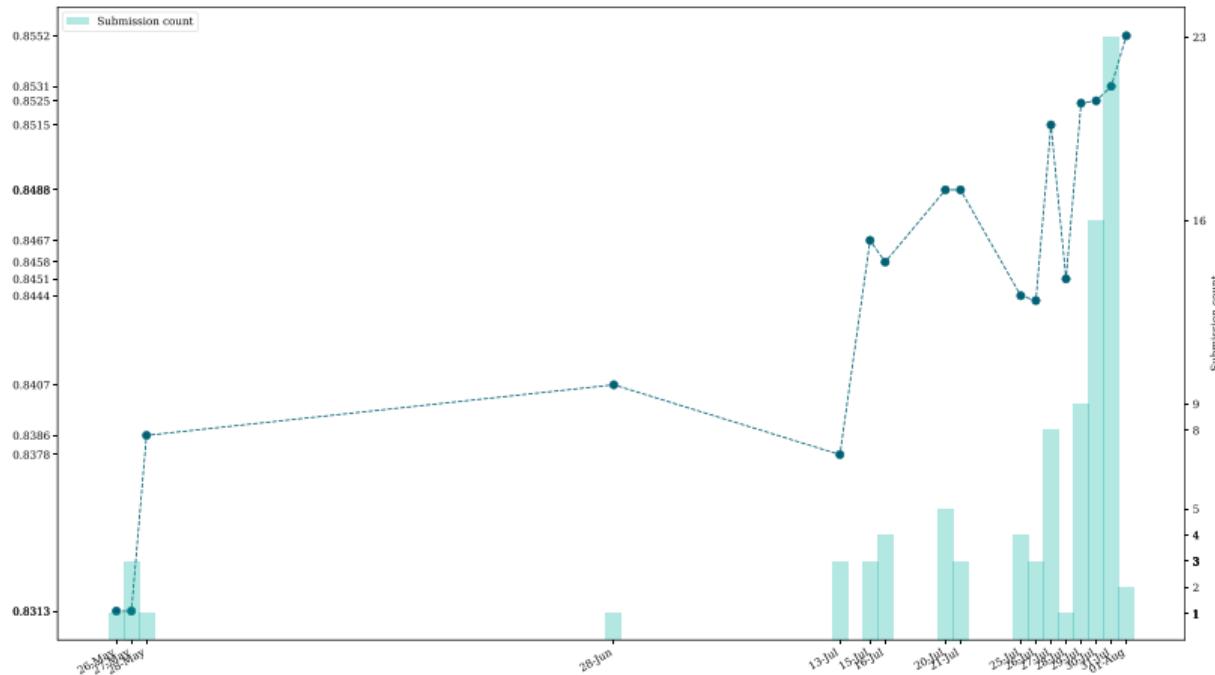
Tuy nhiên, những phần cứng này cũng có giới hạn

- 1x GTX 3080 (laptop): Khá là mạo hiểm khi để treo laptop 24/24
- 2x GTX 2080 (server): Server này đang treo thực nghiệm của khóa luận (khóa luận cử nhân)

fit@hcmus

Tiến độ của nhóm

Submission và score theo thời gian

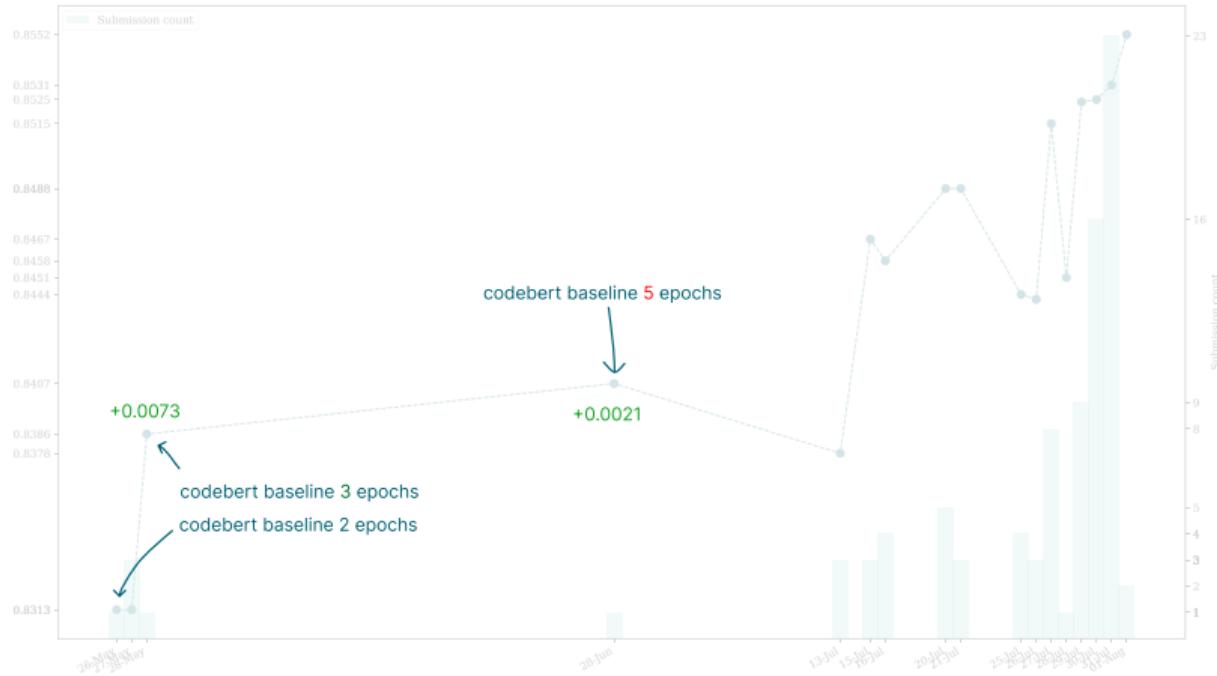


Tiến độ của nhóm

Baseline: dựa trên suicaokhoailang

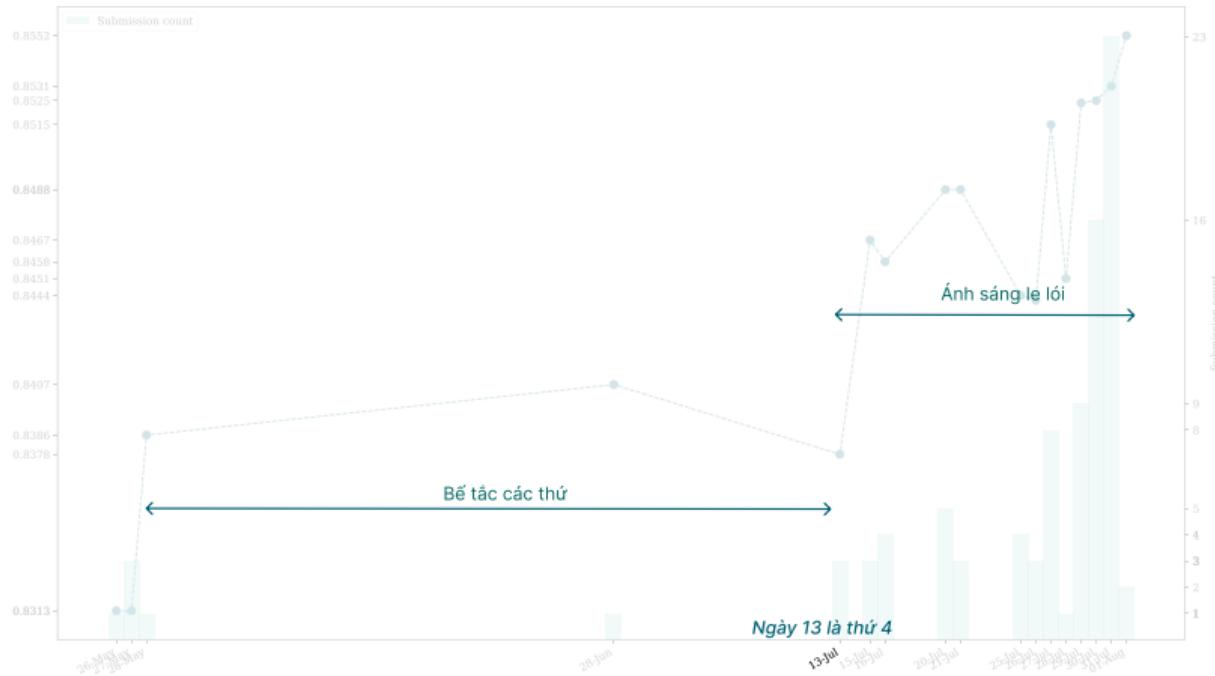
Tiến độ của nhóm

Giai đoạn lung lay niềm tin



Tiến độ của nhóm

Giai đoạn bê tắc



Một số kiến thức tiên nghiệm

Một số kiến thức tiên nghiệm

fit@hcmus

Một số kiến thức tiên nghiệm

Bí nêu hạ quyết tâm đi hỏi

Nhóm hạ quyết tâm và đặt mục tiêu rõ ràng, chỉ cần trên 0.85xx (lúc này là 0.840x). Từ đó nhóm lọc ra các kagglers tiềm năng để đi hỏi trên tinh thần học hỏi. Qua đó, nhóm nhận thấy:

- Ai cũng xài codebert, cha đẻ của codebert cũng có tham gia challenge này và cứ cách 15 ngày lại trồi lên submit.
- Ai cũng có phần cứng khủng, trong 3 kagglers phản hồi lại nhóm thì có một kĩ sư NVIDIA với 8x A100, 2 Data Scientist ở Argentina và Taiwan đều dùng GTX 3090.
- Ai cũng không chỉ dùng single model.

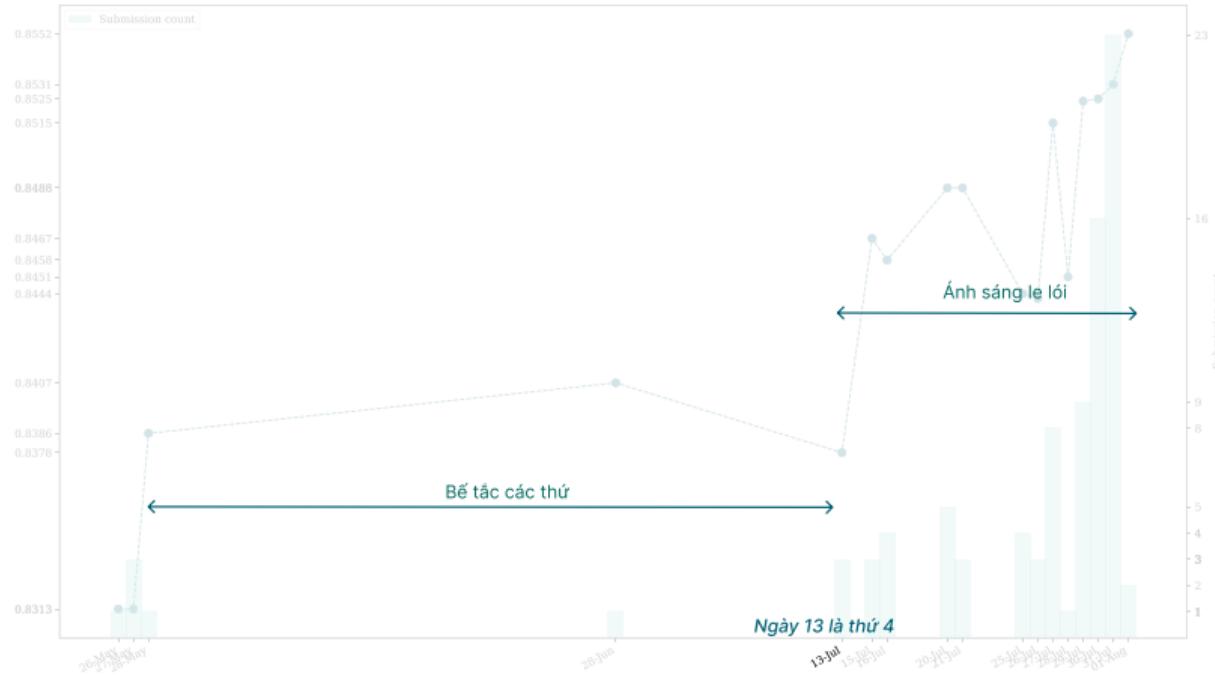
Tới khi chỉ còn 15 ngày cho đến deadline

- Nhóm tiến hành treo 2 mô hình codebert trên laptop GTX 3080 (25 epochs) và GTX 2080 (20 epochs). Baseline này chưa viết lại nên chỉ khi huấn luyện xong thì mới có file checkpoint.
- Nhóm cũng viết lại baseline khác, cho phép saved từng checkpoint trên từng epochs và evaluate từng checkpoint đó.

- Bắt đầu ensemble và thử nhiều các mô hình ensemble khác nhau, từ 1-1, 2-1 cho đến 3-1. Nhóm nhận ra tỉ lệ 3 mô hình pointwise và 1 pairwise là vừa tròn 8:35hrs, đạt hiệu suất tối đa.
- Thử tăng tốc quá trình huấn luyện. Tăng tốc có thể hiểu theo hai hướng:
 - Tăng tốc quá trình huấn luyện trên từng epochs (ví dụ: sử dụng multiple GPU, yêu cầu thêm phần cứng.)
 - Tăng tốc quá trình hội tụ của mô hình trên từng epochs (khả thi).

Tiến độ của nhóm

Giai đoạn có xíu ánh sáng



fit@hcmus

Các mô hình trong ensemble

Dựa trên notebook của thedevastator, nhóm thu được hai mô hình pairwise và 1 mô hình pointwise codebert (tạm gọi là codebert public). Từ đó nhóm tiến hành ensemble thử 4 mô hình như sau:

- codebert 5 epochs
- codebert mlm
- codebert public
- pair-wise

Chiến lược ensemble

Có hai chiến lược như sau:

- Ensemble không thay đổi các vị trí code cell bằng cách ensemble các predict của các mô hình point-wise.
- Ensemble có thay đổi các vị trí code cell bằng cách ensemble các submission rank của từng mô hình dự đoán.

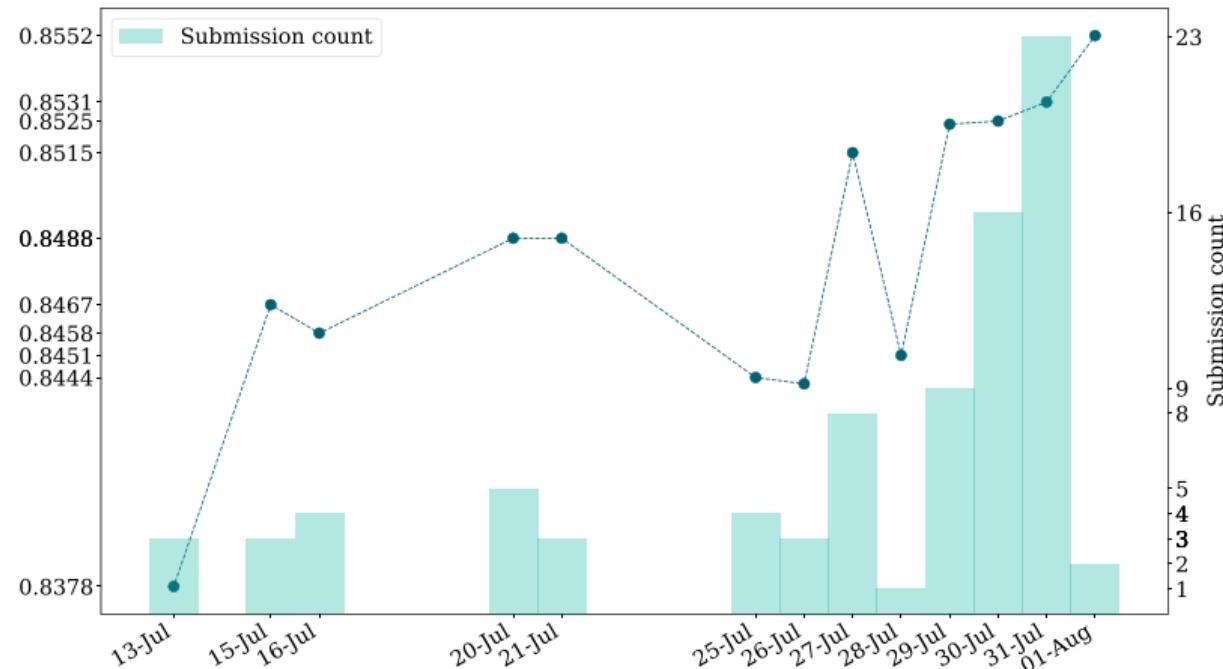
Chiến lược ensemble

Nhìn qua thì có thể thấy chiến lược thứ nhất tốt hơn, tuy nhiên khi test thử thì chiến lược thứ hai lại nhỉnh hơn chiến lược thứ nhất rất nhiều. Lý do có thể giải thích là vì độ đo của bài toán đánh mạnh vào recall hơn precision. Do đó việc ensemble theo chiến lược 2 có xu hướng xê dịch các cell về vị trí tương đối nhiều hơn, trong khi đó chiến lược 1 lại cho ra kết quả chính xác trên từng cell tốt hơn - nhưng tradeoff nhiều hơn cho các cell bị dự đoán sai.

fit@hcmus

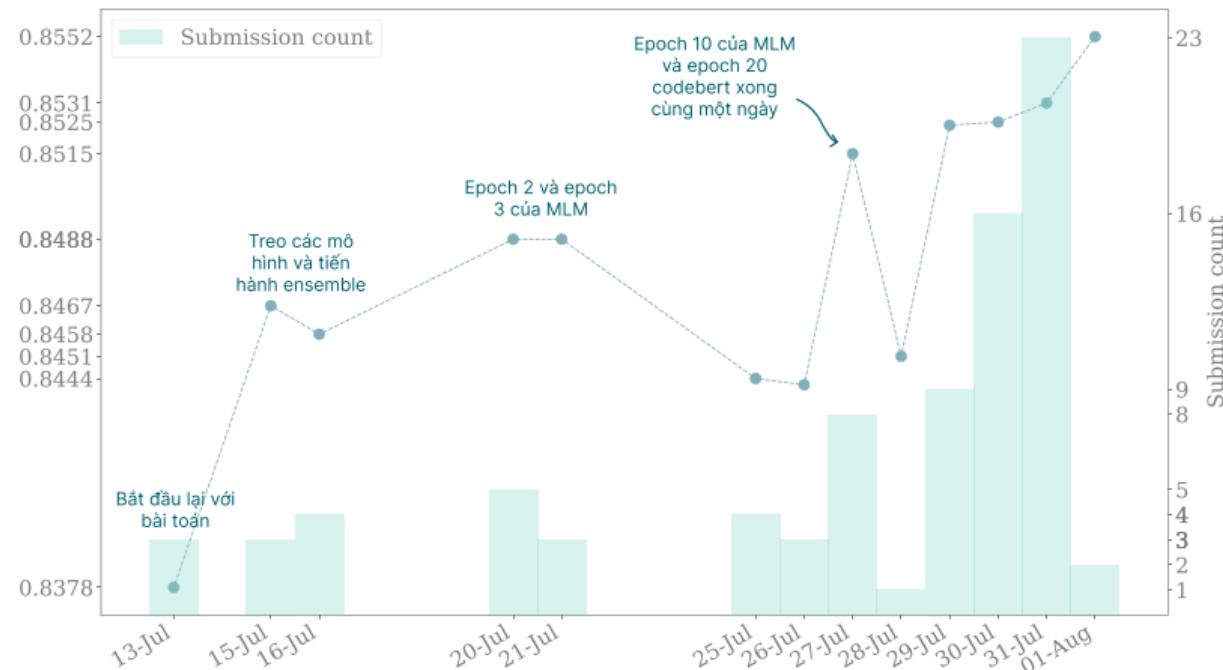
Tiến độ của nhóm

Giai đoạn có xíu ánh sáng



Tiến độ của nhóm

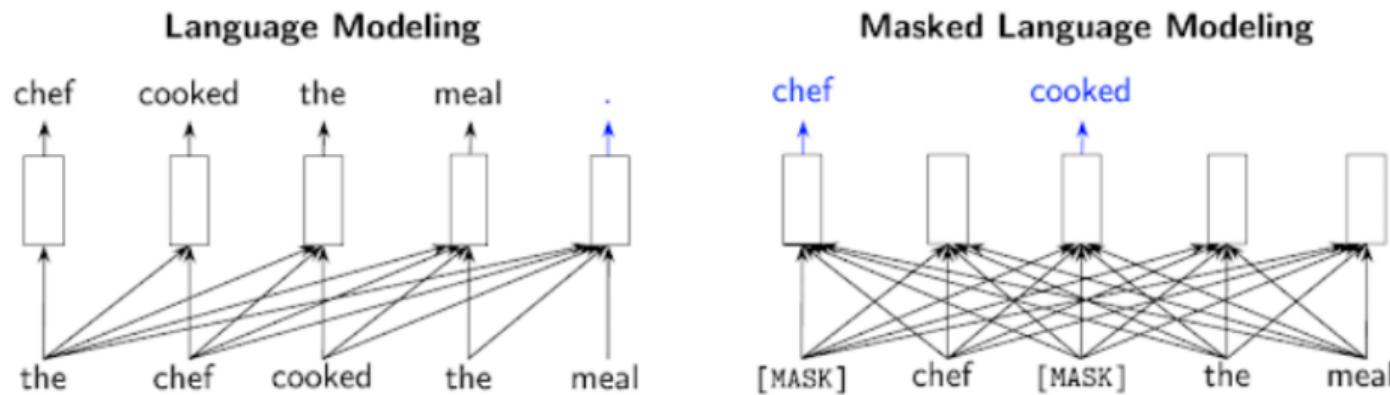
Giai đoạn có xíu ánh sáng



Tiến độ của nhóm

Thử nghiệm MLM

Masked Language Modeling (MLM) là một chiến lược trong fine-tune các mô hình họ BERT.



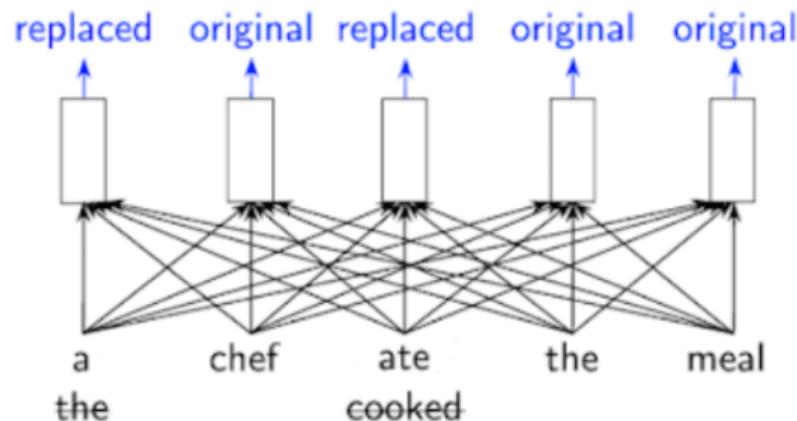
Existing pre-training methods and their disadvantages. Arrows indicate which tokens are used to produce a given output representation (rectangle). Left: Traditional language models (e.g., GPT) only use context to the left of the current word. Right: Masked language models (e.g., BERT) use context from both the left and right, but predict only a small subset of words for each input.

Tiến độ của nhóm

RTD

Replaced Token Detection (RTD) là một chiến lược cải tiến của MLM, được proposed trong bài báo giới thiệu mô hình ELECTRA[1].

Replaced Token Detection



fit@hcmus

Replaced token detection trains a bidirectional model while learning from all input positions.



Tiến độ của nhóm

Codebert

CODEBERT's Objective

Codebert objective là sử dụng mixed giữa MLM và RTD, trong đó MLM đục lỗ trên cả hai đầu dữ liệu NL-PL, và RTD dùng để học các dữ liệu unimodal (dữ liệu không có pair).

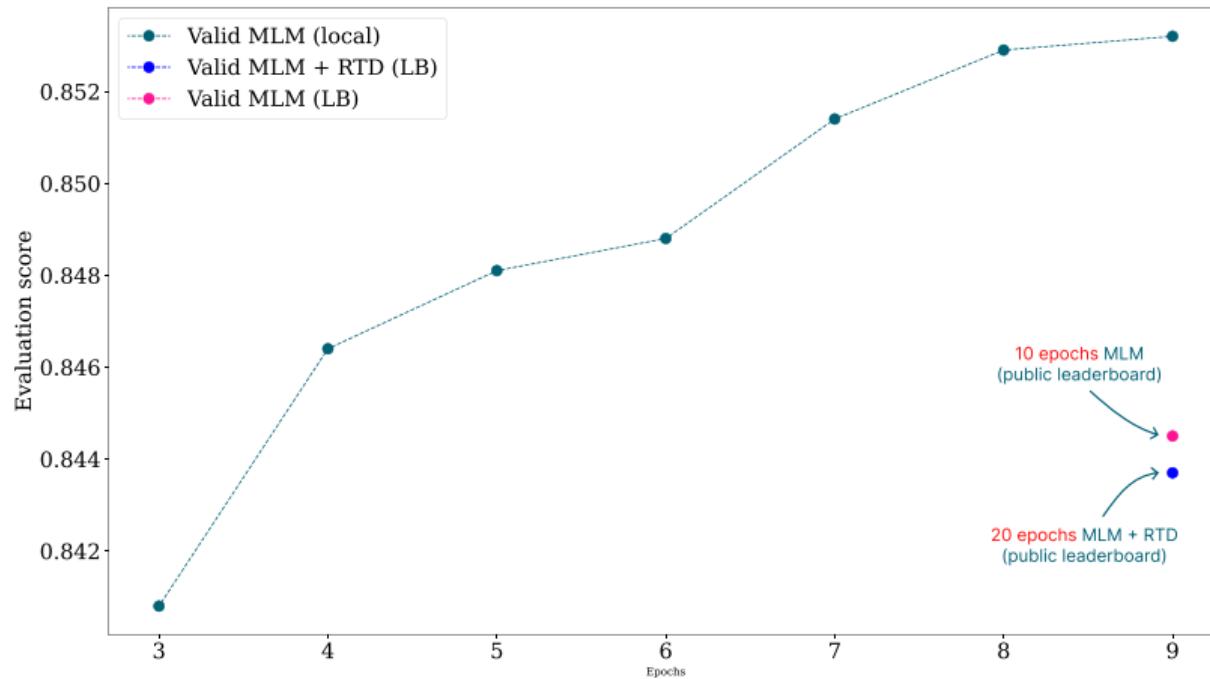
Motivation

Dễ thấy, động lực của RTD là việc tận dụng học trên các dữ liệu unimodal. Tuy nhiên, dữ liệu của chúng ta có là bimodal, trong đó có các pair md-code. Hơn hết, dữ liệu unimodal đã được fine-tune tốt trên codebert. Assumption của nhóm là quá trình mixed giữ RTD và MLM đã khiến mô hình hội tụ lâu hơn thông thường, nên nhóm tiến hành thí nghiệm lược bỏ RTD trong quá trình fine-tune.

fit@hcmus

Tiến độ của nhóm

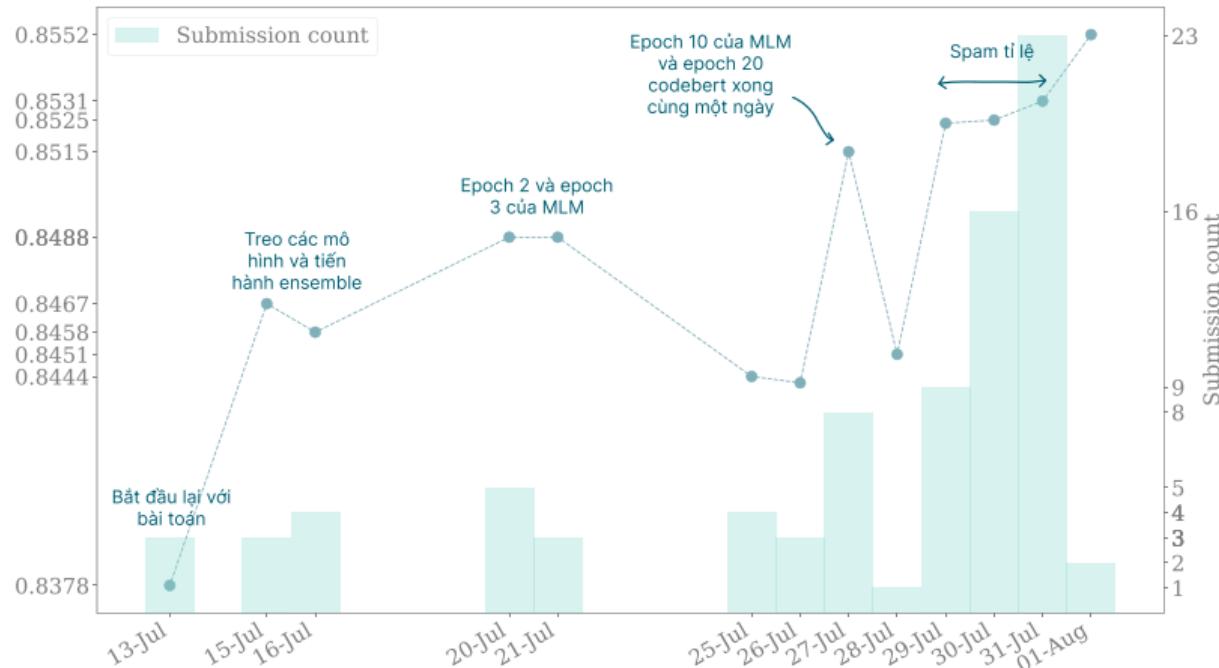
Kết quả thực nghiệm với MLM



Phương pháp của nhóm

Phương pháp của nhóm

Giai đoạn có xíu ánh sáng



Phương pháp của nhóm

Các mô hình ensemble

Các mô hình ensemble bao gồm:

- CODEBERT 20 epochs
- CODEBERT MLM 10 epochs
- CODEBERT public
- Pair-wise

Tỉ lệ ensemble đạt kết quả tốt nhất là 6:5:2:3.

Phương pháp của nhóm

Kết quả sau cùng

Trước 31 tháng 7,

61 HCMUS - Atlantis



0.8531

100

1d



Your Best Entry!

Your submission scored , which is not an improvement of your previous score. Keep trying!

11:30 PM ngày 1/8,

47 HCMUS - Atlantis



0.8552

101

8h



Your Best Entry!

Your most recent submission scored 0.8552, which is an improvement of your previous score of 0.8531. Great job!

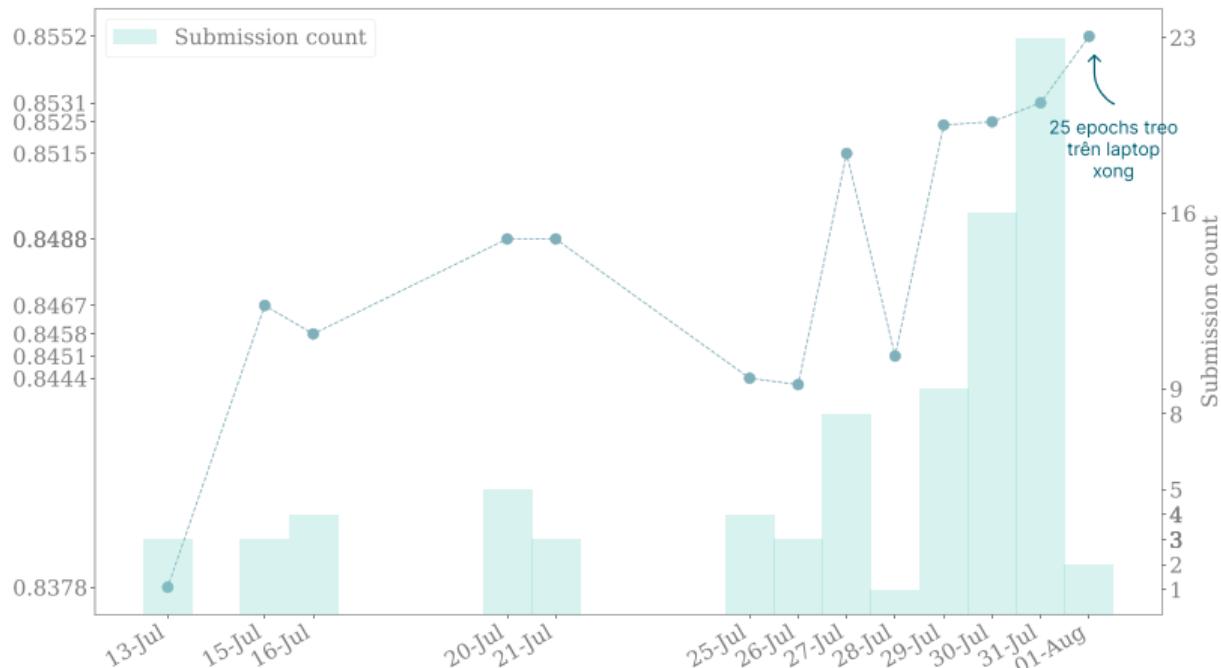
[Tweet this](#)

HUS



Phương pháp của nhóm

Món quà đến muộn



fit@hcmus

Phương pháp của nhóm

Các mô hình ensemble

Các mô hình ensemble bao gồm:

- CODEBERT 20 epochs
- CODEBERT MLM 10 epochs
- CODEBERT 25 epochs
- Pair-wise

Tỉ lệ ensemble đạt kết quả tốt nhất là 6:5:2:3, đạt kết quả 0.8552.

Các hướng phát triển tương lai

Các hướng phát triển tương lai

fit@hcmus

Các hướng phát triển tương lai

Nhóm có thể làm tốt hơn nếu

Giảm thiểu thời gian huấn luyện

- Nếu bạn có nhiều hơn một GPU, bạn có thể sử dụng `torch.distributed` để tăng tốc quá trình huấn luyện mô hình.
- Nếu bạn chỉ có một GPU, bạn có thể hướng tới việc tối ưu trên transformer (nhóm HCMUS-FairyTail).

fit@hcmus

Các hướng phát triển tương lai

Nhóm có thể làm tốt hơn nếu

Force prediction trên các mô hình pair-wise

Các quan hệ của notebook cell như (a.) $>$ (b.), '#' $>$ '##', v.v. Có thể được force vào mô hình pair-wise dưới dạng một prior knowledge để ensemble, giúp mô hình ăn chắc các dự đoán đúng. Nhóm không thử thực nghiệm này được vì dữ liệu của mô hình pair-wise đã được tensor hóa, nhóm không giành thêm thời gian cho hướng này được (vì rất tốn thời gian tìm quan hệ để force).

fit@hcmus

Các hướng phát triển tương lai

Nhóm có thể làm tốt hơn nếu

Chiến lược ensemble tốt hơn

Một chiến lược ensemble tốt hơn là vẫn giữ được thứ tự đúng của các code cell, nhưng vẫn có thể ensemble các markdown cell và giảm các inversion. Nhóm đã cài đặt một phiên bản ensemble này nhưng vì lập trình chưa tới nên inference chậm, chỉ ensemble được 2 mô hình với nhau.

fit@hcmus

Các hướng phát triển tương lai

Nhóm có thể làm tốt hơn nếu

Giảm thời gian inference

Giảm thời gian inference của từng mô hình sẽ giúp stack được nhiều mô hình hơn vào submission.

fit@hcmus

Kết luận

Kết luận

fit@hcmus

Kết luận

- Từ việc đi thảo luận và học hỏi, nhóm đã có được lượng insight giá trị để hoàn thiện project này. Các kỹ thuật học được không chỉ áp dụng được cho bài toán này mà có thể áp dụng lên các bài toán liên quan, nhất là các mô hình họ BERT.
- Challenge này qui tụ rất nhiều những nhân tài, các solution sau giai đoạn public leaderboard hứa hẹn sẽ mang nhiều giá trị cho các bài toán trên miền xử lý dữ liệu ngôn ngữ tự nhiên.

Một số tham khảo

Một số tham khảo

fit@hcmus

Một số tham khảo

- Mã nguồn multiple GPU
<https://github.com/Xianchao-Wu/ai4code-baseline>
- Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators.
- Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., Shou, L., Qin, B., Liu, T., Jiang, D., & Zhou, M. (2020). CodeBERT: A Pre-Trained Model for Programming and Natural Languages.

Lời cảm ơn

Lời cảm ơn

fit@hcmus

Lời cảm ơn

Quá quá trình làm việc, nhóm đã trải qua nhiều lần bế tắc trong việc tiếp tục train codebert hay tìm một cách khác để tăng được kết quả. Nhóm đã thu nhận được rất nhiều insight (nhờ mạnh dạn đi spam hỏi toàn bộ những kagglers hơn mình thông qua email và Linkedin) từ các Kaggler khác, trong đó xin chân thành cảm ơn các phản hồi giá trị:

- Xianchao Wu, một PhD và một kỹ sư AI cho NVIDIA Nhật Bản. Nhóm đã có được insight có thể thực hiện multi-gpu với torch-distributed. Một kỹ thuật đòi hỏi điều kiện phần cứng cao nên đã không thực hiện được trong đồ án này, nhưng việc học hỏi được một kỹ thuật tăng tốc như vậy sẽ giúp nhóm trong chặng đường nghiên cứu lâu dài phía trước.
- Sergio Manuel Papadakis, một kỹ sư AI và Hạt Nhân ở Argentina. Nhóm đã có được insight của phương pháp ensemble và tối ưu thời gian ensemble cho từng mô hình thông qua discuss với kaggle này.
- Mark Peng, một Data Scientist ở Taiwan. Nhóm đã có được insight của việc thay đổi objective.

fit@hcmus

Bonus

Bonus

fit@hcmus

Phương pháp của nhóm

Kết quả gần đây

Kết quả có được sau khi ensemble với mô hình graph-codebert:

63 HCMUS - Atlantis



0.8558

109

21h

fit@hcmus

Question and answer

fit@hcmus

Phụ lục

fit@hcmus

Slide's color palette

PANTONE 17-3938 Very Peri



Figure 4: The Pantone Color of the Year 2022 ¹

fit@hcmus

¹<https://www.pantone.com/color-of-the-year-2022>

Thanh you :3

fit@hcmus