ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP LỚN
MÔN HỌC MÁY - MACHINE LEARNING
Ngành: Khoa học máy tính

Nguyễn Văn Huy Hoàng
Nguyễn Viết Thành Lân
Nguyễn Văn Thanh Tùng

HÀ NỘI - 2025

ĐẠI HỌC QUỐC GIA HÀ NỘI
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP LỚN
MÔN HỌC MÁY - MACHINE LEARNING

Ngành: Khoa học máy tính

Nguyễn Văn Huy Hoàng - 23021561
Nguyễn Viết Thành Lân - 23020099
Nguyễn Văn Thanh Tùng - 23021716

Cán bộ hướng dẫn: PhD. Tạ Việt Cường
MsC. Lê Bàng Giang

HÀ NỘI - 2025

BÁO CÁO BÀI TẬP LỚN MÔN HỌC MÁY
CAFA 6 PROTEIN FUNCTION PREDICTION

Sinh viên báo cáo

Nguyễn Văn Huy Hoàng
K68I – CS3
23021561@vnu.edu.vn

Nguyễn Viết Thành Lân
K68 – IT3
23020099@vnu.edu.vn

Nguyễn Văn Thanh Tùng
K68I – CS2
23021716@vnu.edu.vn

Công việc

Tên công việc	Tên thành viên phụ trách
Xử lý dữ liệu	Nguyễn Văn Huy Hoàng, Nguyễn Văn Thanh Tùng
Đề xuất, xây dựng mô hình	Nguyễn Văn Huy Hoàng, Nguyễn Viết Thành Lân
Huấn luyện mô hình	Nguyễn Viết Thành Lân, Nguyễn Văn Thanh Tùng
Tổng hợp báo cáo và kết quả	Nguyễn Văn Huy Hoàng, Nguyễn Viết Thành Lân, Nguyễn Văn Thanh Tùng

Tóm tắt

CAFA 6 Protein Function Prediction là cuộc thi tổ chức trên Kaggle về huấn luyện mô hình học máy để dự đoán các thuật ngữ về Gene Ontology (GO) cho tập hợp các protein dựa trên trình tự axit amin của chúng. Bài báo cáo này trình bày về hiệu quả của các kỹ thuật học máy được sử dụng để giải quyết và huấn luyện mô hình cho cuộc thi. Nhóm chỉ sử dụng dữ liệu do cuộc thi cung cấp và kết quả của mô hình huấn luyện cũng được đánh giá theo chuẩn của Kaggle. Cuối cùng đưa ra được những giải thích và kết luận về tỉ lệ chính xác của công việc.

(Keywords - Machine Learning,

liệu pháp y tế mới.

Nhóm chú trọng đến vấn đề giải

I. Giới thiệu

Về cuộc thi CAFA 6 Protein Function Prediction: Protein là những phân tử lớn chịu trách nhiệm cho nhiều hoạt động trong tế bào, mô, cơ quan và cơ thể chúng ta, đồng thời đóng vai trò trung tâm trong cấu trúc và chức năng của tế bào. Protein được cấu tạo từ 20 loại phân tử nhỏ hơn gọi là axit amin, được sắp xếp thành một chuỗi dài gọi là trình tự axit amin của protein. Mỗi protein có trình tự riêng quyết định cấu trúc và chức năng của nó. Bạn sẽ xây dựng một mô hình dự đoán chức năng của protein dựa trên trình tự axit amin của nó. Những dự đoán này sẽ giúp các nhà nghiên cứu hiểu cách thức và

Tiếp theo là phân tích và xây dựng phân cấp Gene Ontology (GO Hierarchy).

quyết bài toán do cuộc thi đưa ra, tìm kiếm một mô hình hiệu quả cho dự đoán chức năng protein. Đoạn code là một quy trình hoàn chỉnh đi từ xử lý đầu vào đến embeddings, sau đó xây dựng mô hình dự đoán và kiểm thử trên mẫu.

II. Phương pháp

Phương pháp của nhóm là xây dựng mô hình dựa trên các phương pháp học máy, học sâu cơ bản. Các phương pháp sẽ được giải thích rõ theo từng mục sau đây:

2.1. Nguồn dữ liệu và xử lý dữ liệu

Nguồn dữ liệu gốc đã được cung cấp trên Kaggle, bao gồm 8 tệp (4 tệp huấn luyện, 2 tệp kiểm tra, 1 tệp bài nộp mẫu và 1 tệp trọng số) và các tệp dữ liệu đã được nhóm xử lý embeddings:
"https://www.kaggle.com/competitions/cafa-6-protein-function-prediction/data"

Đầu tiên là xác định nền tảng, nhóm đã xử lý tiền dữ liệu bằng phương pháp thay thế dữ liệu. Sau thống kê, nhận thấy tất cả dữ liệu đều không bị bỏ trống, song có những dữ liệu nhiễu bị sai, không thuộc 20 ký tự của Axit Amin. Nhóm đã quyết định xử lý bằng cách thay các ký tự lạ không thuộc 20 ký hiệu bằng ký tự dự phòng "X". Đối với các trình tự protein có nhiều ký tự lỗi với ngưỡng *invalid_threshold = 0.5* sẽ được loại bỏ. Sau bước xử lý dữ liệu thô bên trên, nhóm tiếp tục phân loại và xử lý nhãn. Các nhãn GO sẽ được phân loại theo từng EntryID để thiết lập mối quan hệ Protein -Term.

Từ xử lý tiền dữ liệu, nhóm sử dụng các mô hình embeddings để cho ra các file .npy nhằm phục vụ cho mô hình dự đoán sau này.

Sau huấn luyện, các đầu vào được đánh giá qua ba mô hình độc lập, đưa ra Nhóm xây dựng lên đồ thị quan hệ cha con GO graph, mục đích nhằm xử lý tính nhất quán lan truyền tổ tiên và đảm bảo tính nhất quán sinh học của kết quả dự đoán. Cuối cùng, để tinh chỉnh dữ liệu nhãn nhóm sử dụng thêm chú thích ngoài GOA. Trong đó các chú thích mang định tính NOT được lan truyền âm tính xuống toàn bộ các thuật ngữ hậu duệ, hình thành tập ràng buộc negative_keys nhằm ngăn mô hình đưa ra các dự đoán sai về mặt sinh học, đồng thời các nhãn dương tính đáng tin cậy được giữ lại làm ground truth cho giai đoạn hợp nhất cuối.

2.2. Huấn luyện mô hình học sâu đa nhãn

Phần này, nhóm chia thành ba mô hình cho ba bài toán con. Mỗi bài toán con là một khía cạnh của GO - Chức năng Phân tử (MFO), Quá trình Sinh học (BPO) và Thành phần Tế bào (CCO). Danh sách các Go term được chuyển đổi thành ma trận nhị phân thưa thớt và từng khía cạnh được mã hóa bằng MultiLabelBinarizer khác nhau và xác định số lớp cụ thể. Nhóm em chọn mô hình mạng nơ-ron đa lớp Simple MLP cho huấn luyện. Mô hình mạng còn kết hợp Batch Normalization, ReLU và Dropout nhằm đảm bảo khả năng biểu diễn mạnh trong khi vẫn kiểm soát overfitting. Quá trình huấn luyện được sử dụng hàm mất mát BCE kết hợp Sigmoid và Binary Cross-Entropy. Bộ tối ưu AdamW cùng cơ chế điều chỉnh learning rate nhằm thích ứng, có thể giảm learning rate nếu lỗi validation ngừng cải thiện.

2.3. Dự đoán đa mô hình

Function (F) và Biological Process (P).

các logits. Các logits đó được tính toán theo hàm Sigmoid và chuyển thành xác suất dự đoán P. Các kết quả sẽ được lựa chọn ngưỡng kép theo hai tiêu chí, P cao hơn 0.2 và top 25 P cao nhất mới được giữ lại. Cuối cùng giai đoạn hậu xử lý, các dự đoán lan truyền lên tổ tiên của chúng trong cây GO và các cặp protein vi phạm ràng buộc âm tính sẽ được loại bỏ, còn các chú thích GOA tin cậy thì được tích hợp trở lại với xác suất 1.0.

III. Đánh giá

3.1. Mô hình và huấn luyện

Trong dự án này, nhóm em sử dụng Multi-Layer Perceptron (MLP) để giải quyết bài toán phân loại đa nhãn giữ trên vector nhúng của protein. Đoạn code sử dụng mô hình MLP 3 lớp (3-layer MLP) hoạt động trên nền tảng vector đặc trưng được trích xuất từ ESM-2:

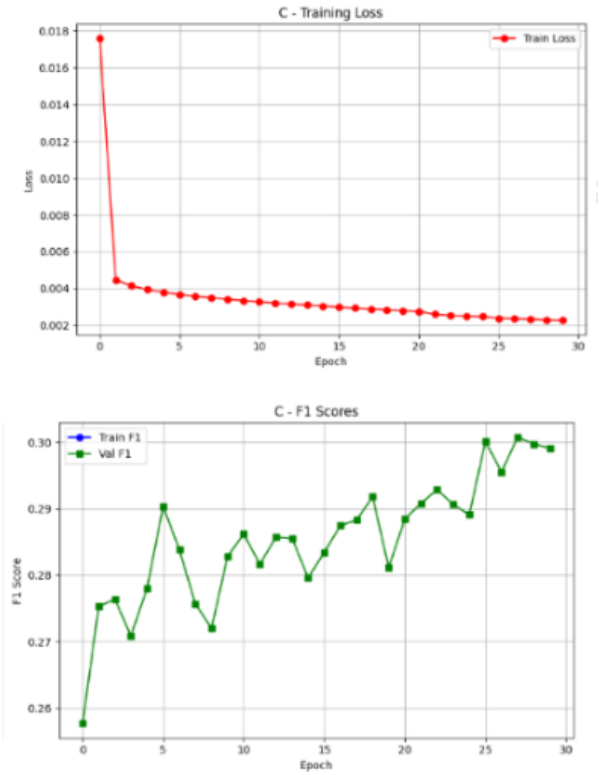
- Input layer: 1280 chiều
- Hidden layer 1: Nhận vector đầu vào 1280 chiều rồi giảm xuống 1280 chiều, kết hợp chuẩn hóa và loại bỏ bớt số chiều để tránh học vẹt.
- Hidden layer 2: Tiếp tục giảm từ 1024 xuống 512 chiều.
- Output layer: Dựa trên 512 đặc điểm này để xác định protein có bao nhiêu phần trăm thuộc chức năng nào

3.2. Kết quả huấn luyện theo từng khía cạnh

Việc đánh giá được thực hiện riêng biệt trên 3 khía cạnh chính của GO: Cellular Component (C), Molecular

3.2.1. Cellular Component (C)

Dữ liệu: Number of proteins: 60292, GO terms: 2651.



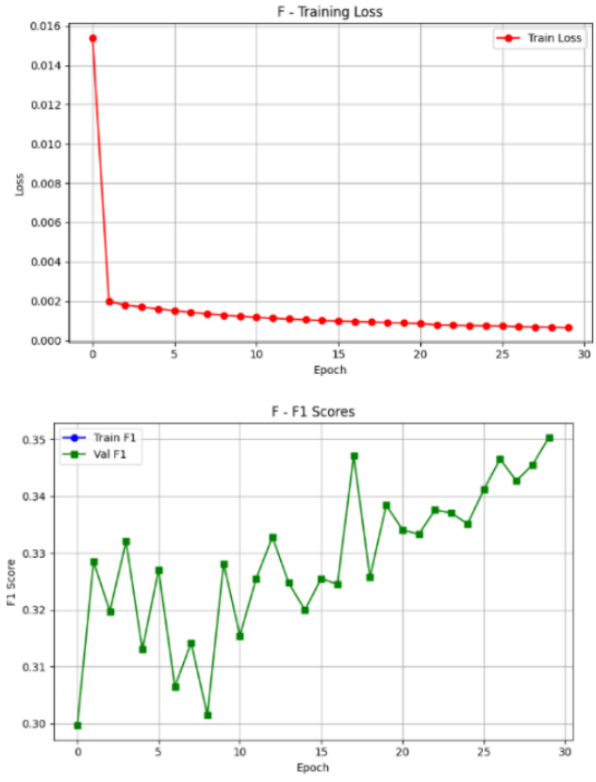
H3.1. Biểu đồ Cellular Component (C)

Nhận xét:

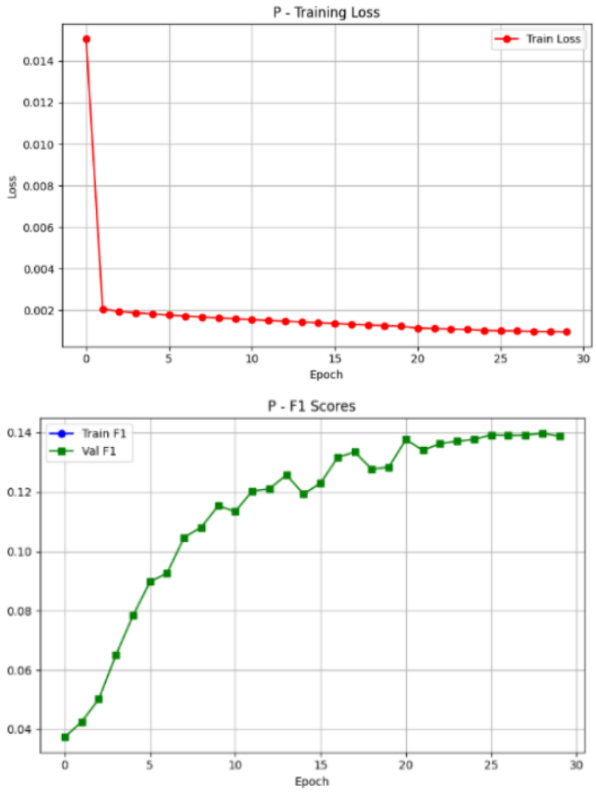
- Training Loss: giảm nhanh xuống mức thấp (gần 0.002) và dần ổn định sau epoch 25.
- F1 Scores: điểm F1 có xu hướng tăng theo thời gian và đạt đỉnh (0.3) vào khoảng epoch 25 đến 30.Điều này cho thấy mô hình học khá tốt các đặc trưng về thành phần tế bào.

3.2.2. Molecular Function (F)

Dữ liệu: Number of proteins: 58001, GO terms: 6616.



H3.2. Biểu đồ Molecular Function (F)



H3.3. Biểu đồ Biological Process (P)

Nhận xét:

- Training Loss: giảm nhanh xuống mức thấp (dưới 0.001) và dần ổn định sau epoch 20.
- F1 Scores: điểm F1 giao động rất mạnh trong khoảng 0.3-0.35.Sự không ổn định này có thể do tính chất phân bố dữ liệu của các chức năng phân tử phức tạp hơn.

3.2.3 Biological Process (P)

Dữ liệu: Number of proteins: 59958, GO terms: 16858.

Nhận xét:

- Training Loss: giảm nhanh xuống mức thấp tương tự khía cạnh F
- F1 Scores: điểm F1 tăng trưởng đều từ 0.04 và đạt đỉnh ở 0.14. Mặc dù điểm số thấp hơn so với C và F, nhưng đường cong F1 mượt mà (ít dao động) cho thấy mô hình đang học ổn định, tuy nhiên kiến trúc hiện tại có thể đã chạm giới hạn khả năng học cho khía cạnh phức tạp này.

3.3. Tổng kết

Kết quả thực nghiệm cho thấy kiến trúc MLP kết hợp với ESM embeddings hoạt động hiệu quả nhất trên khía cạnh Molecular Function (F) và Cellular Component (C). Đối với Biological Process (P), do không gian nhãn quá lớn, mô hình gặp khó khăn trong việc đạt độ chính xác cao.

IV. Kết luận

Trong báo cáo này, nhóm đã xây dựng mô hình dự đoán chức năng protein CAFA 6 Protein Function Prediction. Kết quả tốt nhất mà nhóm đạt được trong quá trình xây dựng mô hình là 0.278 theo đánh giá của Kaggle.

Dưới đây là link code của nhóm: https://github.com/NguyenVanThanhTung/Protein_INT3405E_2

Dưới đây là link data của nhóm:

1. Dữ liệu protein embeddings: Dữ liệu 01
2. Dữ liệu go annotation: Dữ liệu 02

Tài liệu tham khảo

1. Tham khảo tại: Tham khảo