ĐẠI HỌC QUỐC GIA TPHCM

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

AI23 (23TNT1), FIT@HCMUS-VNUHCM

Tóm tắt Final Project

Đề tài: CLIP (Contrastive Language-Image Pretraining)

Môn học: Phương pháp Toán cho Trí tuệ nhân tạo

Sinh viên thực hiện:
Nguyễn Đình Hà Dương (23122002)
Nguyễn Lê Hoàng Trung (23122004)
Đinh Đức Tài (23122013)
Hoàng Minh Trung (23122014)

Giáo viên hướng dẫn:
TS. Cấn Trần Thành Trung
ThS. Nguyễn Ngọc Toàn

Ngày 20 tháng 5 năm 2025

1 Tóm tắt

Chúng tôi lựa chọn **Hướng 2: Nghiên cứu một bài báo khoa học** với lý do sau:

- Nâng cao hiểu biết về các mô hình cơ bản, quan trọng trong lĩnh vực Machine learning / Deep learning / AI.
- Hiểu cách ứng dụng của toán học và kĩ thuật lập trình trong nghiên cứu và thực tế.
- Tăng khả năng tự nghiên cứu và áp dụng tri thức mới vào thực tiễn.
- Nâng cao tư duy phản biện, phân tích, tổng hợp.

CLIP [1] (Contrastive Language-Image Pre-Training) là một mạng nơ-ron được huấn luyện trên nhiều cặp dữ liệu (ảnh, văn bản). Mô hình này có thể được hướng dẫn bằng ngôn ngữ tự nhiên để dự đoán đoạn văn bản phù hợp nhất với một hình ảnh, mà không cần được tối ưu hóa trực tiếp cho nhiệm vụ đó — tương tự như khả năng "zero-shot" (không cần huấn luyện lại) của GPT-2 và GPT-3.

CLIP đạt hiệu suất tương đương với ResNet-50 gốc trên bộ dữ liệu ImageNet theo cách "zero-shot", mà không cần sử dụng bất kỳ ví dụ có gán nhãn nào trong 1,28 triệu ảnh ban đầu, từ đó vượt qua một số thách thức lớn trong lĩnh vực thị giác máy tính.

CLIP [1], Code [2], Blog [3].

Lý do lựa chọn bài báo:

- CLIP là một mô hình cơ sở (foundation model) quan trọng trong lĩnh vực học đa phương thức, kết nối giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên.
- CLIP có ứng dụng đa dạng: truy vấn hình ảnh bằng văn bản; phân loại hình ảnh zero-shot và fewshot; phân tích nội dung video.
- CLIP được xây dựng dựa trên các kiến thức nền tảng toán học như đại số tuyến tính, xác suất thống kê. Hiểu rõ CLIP sẽ nâng cao kiến thức về toán học.

Kế hoạch thực hiện:

1. Phân tích các nền tảng của CLIP: Contrastive learning, Vision Transformer/Transformer, Embedding Vector Space,...
2. Phân tích dữ liệu, cách huấn luyện CLIP, so sánh các mô hình.
3. Các mô hình SOTA được phát triển thêm dựa trên CLIP: BLIP-2, BLIP3-o,...
4. Ứng dụng của CLIP vào cuộc thi truy vấn AI Challenge HCMC 2025.

Tài liệu

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[2] OpenAI. CLIP: Contrastive Language-Image Pre-training (Software Repository). https://github.com/openai/CLIP, 2021. Accessed: 2025-05-20.

[3] OpenAI. CLIP: Connecting text and images. https://openai.com/index/clip/, 2021. Accessed: 2025-05-20.