THÔNG TIN CHUNG CỦA NHÓM

- Link YouTube video của báo cáo (tối đa 5 phút): *https://youtu.be/rGWEgQVjfO4*
- Link slides (dạng .pdf đặt trên Github của nhóm): CS2205.CH201/Hồ Hồ Đặng Thanh - CS2205.SEP2025.DeCuong.FinalReport.Slide.pdf at main · Thanh-Hoo/CS2205.CH201
- *Mỗi thành viên của nhóm điền thông tin vào một dòng theo mẫu bên dưới*
- *Sau đó điền vào Đề cương nghiên cứu (tối đa 5 trang), rồi chọn Turn in*
- *Lớp Cao học, mỗi nhóm một thành viên*

- Họ và Tên: Hồ Đặng Thanh Hồ
- MSSV: 250101021
- Lớp: CS2205.CH201
- Tự đánh giá (điểm tổng kết môn): 9/10
- Số buổi vắng: 0
- Số câu hỏi QT cá nhân: 8
- Số câu hỏi QT của cả nhóm: 8
- Link Github: https://github.com/Thanh-Hoo/CS2205.CH201

ĐỀ CƯƠNG NGHIÊN CỨU

TÊN ĐỀ TÀI (IN HOA)

NGHIÊN CỨU VÀ ỨNG DỤNG KỸ THUẬT THÍCH NGHI THAM SỐ HIỆU QUẢ (LORA) TRÊN MÔ HÌNH NGÔN NGỮ-THỊ GIÁC CHO BÀI TOÁN PHÁT HIỆN BẤT THƯỜNG TRONG CÔNG NGHIỆP.

TÊN ĐỀ TÀI TIẾNG ANH (IN HOA)

RESEARCH AND APPLICATION OF PARAMETER-EFFICIENT ADAPTATION (LORA) ON VISION-LANGUAGE MODELS FOR INDUSTRIAL ANOMALY DETECTION

TÓM TẮT *(Tối đa 400 từ)*

Trong dây chuyền sản xuất tự động hóa, việc phát hiện sản phẩm lỗi (Anomaly Detection) là bài toán then chốt. Tuy nhiên, các phương pháp Deep Learning giám sát truyền thống đòi hỏi số lượng lớn dữ liệu lỗi được gán nhãn, điều này thường bất khả thi trong thực tế do lỗi xuất hiện rất hiếm. Gần đây, các mô hình nền tảng Ngôn ngữ-Thị giác (Vision-Language Models - VLMs) như CLIP đã cho thấy khả năng nhận diện vật thể vượt trội ở dạng Zero-shot. Mặc dù vậy, CLIP gặp khó khăn khi nhận diện các lỗi chi tiết nhỏ do sự khác biệt về miền dữ liệu (domain gap) giữa ảnh tự nhiên và ảnh công nghiệp.

Khoá luận này đề xuất phương pháp tinh chỉnh mô hình CLIP sử dụng kỹ thuật thích nghi hạng thấp (Low-Rank Adaptation - LoRA) kết hợp với kỹ thuật học prompt (Prompt Learning) trong bối cảnh Few-shot (học từ vài mẫu). Phương pháp này cho phép mô hình thích nghi nhanh chóng với từng loại sản phẩm cụ thể mà chỉ tốn tài nguyên tính toán rất thấp, không cần huấn luyện lại toàn bộ mạng. Kết quả thực nghiệm sẽ được đánh giá trên bộ dữ liệu chuẩn MVTec AD, so sánh với các phương pháp State-of-the-Art hiện tại như WinCLIP.

GIỚI THIỆU *(Tối đa 1 trang A4)*

1. Bối cảnh và tính cấp thiết

   Sự bùng nổ của Công nghiệp 4.0 đòi hỏi các hệ thống kiểm định chất lượng quang học (AOI) phải hoạt động chính xác và linh hoạt.

   Các mô hình học sâu hiện tại (CNNs, Autoencoders) hoạt động tốt nhưng "đóng kín", nghĩa là khi chuyển sang dây chuyền sản xuất vật thể mới, ta phải thu thập lại dữ liệu và huấn luyện lại từ đầu (Train from scratch), gây tốn kém thời gian và chi phí.

   Sự ra đời của Foundation Models mở ra hướng đi mới: Tận dụng tri thức có sẵn để nhận diện lỗi mà không cần nhiều dữ liệu training.

2. Vấn đề nghiên cứu (Problem Statement)

   Vấn đề 1: Mô hình CLIP gốc được huấn luyện trên dữ liệu internet, nên nó hiểu "cái chai" là gì, nhưng không hiểu "cái chai bị xước dăm" trông như thế nào.

   Vấn đề 2: Fine-tuning toàn bộ mô hình CLIP (hàng trăm triệu tham số) đòi hỏi GPU rất mạnh và dễ dẫn đến hiện tượng "quên tri thức cũ" (Catastrophic Forgetting).

   Câu hỏi nghiên cứu: Làm thế nào để thích nghi mô hình CLIP vào miền dữ liệu công nghiệp với chi phí tính toán thấp nhất mà vẫn đạt độ chính xác cao trong điều kiện dữ liệu hạn chế (Few-shot)?

3. Phạm vi nghiên cứu

   Đối tượng: Hình ảnh sản phẩm công nghiệp 2D (không xét đến video hay 3D point cloud).

   Dữ liệu: Tập trung vào bộ dữ liệu chuẩn **MVTec AD** (gồm 15 loại vật thể như thảm, gỗ, chai, transistor...).

Công nghệ: Mô hình CLIP (ViT backbone), kỹ thuật LoRA, PyTorch Framework.

MỤC TIÊU *(Viết trong vòng 3 mục tiêu)*

Nghiên cứu lý thuyết: Tìm hiểu sâu về kiến trúc Vision Transformers (ViT), mô hình CLIP và cơ chế hoạt động của Low-Rank Adaptation (LoRA).

Đề xuất giải pháp: Xây dựng một pipeline tích hợp LoRA vào khối Visual Encoder của CLIP, kết hợp với cơ chế so sánh Text-Image để phát hiện bất thường.

Thực nghiệm và Đánh giá: Cài đặt thuật toán, huấn luyện trên Google Colab và so sánh hiệu năng (AUROC) với các phương pháp WinCLIP, AnomalyCLIP.

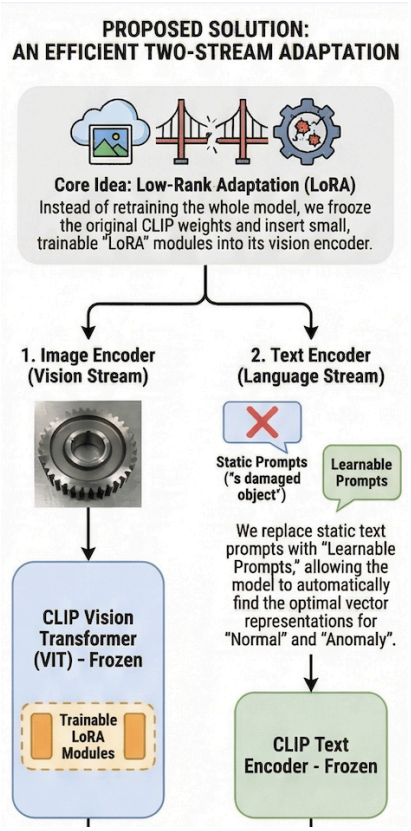
Tối ưu hóa: Chứng minh được phương pháp đề xuất sử dụng ít tài nguyên GPU và thời gian huấn luyện hơn so với Fine-tuning truyền thống.

NỘI DUNG VÀ PHƯƠNG PHÁP

1. Kiến trúc đề xuất (Proposed Architecture)

   Hệ thống sẽ bao gồm 2 nhánh chính (Two-stream network):

   - Image Encoder (Nhánh hình ảnh): Sử dụng CLIP ViT-B/16 đã pre-train.
     - Các trọng số gốc của CLIP sẽ bị đóng băng (Frozen).
     - Các module LoRA được chèn vào các lớp Multi-head Self-Attention. Chỉ có các tham số LoRA này được cập nhật trong quá trình train.
   - Text Encoder (Nhánh văn bản):
     - Thay vì sử dụng prompt thủ công ("A damaged object"), sử dụng kỹ thuật Learnable Prompts (CoOp) để mô hình tự học các vector từ ngữ tối ưu đại diện cho khái niệm "Lỗi" (Anomaly) và "Bình thường" (Normal).



Hình 1. Ý tưởng xây dựng mô hình

2. Quy trình thực hiện (Workflow)
    - **Giai đoạn chuẩn bị:** Tiền xử lý dữ liệu MVTec AD (resize, normalize). Chia tập dữ liệu theo kịch bản Few-shot (ví dụ: chỉ lấy 4, 8, 16 ảnh bình thường để train).
    - **Giai đoạn huấn luyện (Adaptation):**
        - Đưa ảnh và prompt vào mô hình.
        - Tính toán sự tương đồng (Cosine Similarity) giữa đặc trưng ảnh và đặc trưng văn bản.
        - Sử dụng hàm mất mát (Loss Function) như **Focal Loss** hoặc **Cross-Entropy** để tối ưu hóa các trọng số LoRA sao cho khoảng cách giữa ảnh lỗi và prompt "Lỗi" gần nhau nhất.
    - **Giai đoạn kiểm thử (Inference):**
        - Tính toán bản đồ bất thường (Anomaly Map) và điểm số bất

thường (Anomaly Score).

- Phân ngưỡng (Thresholding) để quyết định ảnh là OK hay NG (Not Good).

KẾT QUẢ MONG ĐỢI

1. **Về mặt khoa học:**
   - Đề xuất được một biến thể kiến trúc kết hợp hiệu quả giữa CLIP và LoRA cho bài toán Anomaly Detection.
   - Báo cáo phân tích (Ablation Study) chỉ ra ảnh hưởng của số lượng mẫu (shot) đến độ chính xác.
2. **Về mặt thực tiễn:**
   - Mã nguồn (Source Code) hoàn chỉnh trên Python/PyTorch.
   - Mô hình đạt chỉ số AUROC > 90% (Image-level) trên dataset MVTec AD.
   - Chứng minh khả năng chạy trên GPU phổ thông (như T4 hoặc RTX 3060) với thời gian huấn luyện dưới 30 phút/class.

TÀI LIỆU THAM KHẢO

[1]. **Radford, A., et al. (2021).** *Learning Transferable Visual Models From Natural Language Supervision.* (CLIP Paper).

[2]. **Hu, E. J., et al. (2022).** *LoRA: Low-Rank Adaptation of Large Language Models.* (ICLR 2022).

[3]. **Jeong, J., et al. (2023).** *WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation.* (CVPR 2023).

[4]. **Zhou, Q., et al. (2024).** *AnomalyCLIP: Object-agnostic Prompt Learning for Zero-shot Anomaly Detection.* (ICLR 2024).

[5]. **Bergmann, P., et al. (2019).** *MVTec AD – A Comprehensive Real-World Dataset for Unsupervised Anomaly Detection.* (CVPR 2019).