

**BỘ CÔNG THƯƠNG**

**TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP.HCM**

**KHOA CÔNG NGHỆ THÔNG TIN**

-----🙞🙜🕮🙞🙜-----

**THỊ GIÁC MÁY TÍNH VÀ ỨNG DỤNG**

**TEXTS AS IMAGES IN PROMPT TUNING FOR**

**MULTI-LABEL IMAGE RECOGNITION**

**GVHD: TS. NGUYỄN CHÍ KIÊN**

**Họ tên Sinh Viên MSSV Lớp**

|  |  |  |
| --- | --- | --- |
| **Đoàn Ánh Dương** | **20082631** | **DHKHMT16B** |
| **Nguyễn Hữu Hải** | **20053571** | **DHKHMT16A** |
| **Nguyễn Thị Hồng Vân** | **20003445** | **DHKHMT16A** |
| **Trần Thanh Đức** | **20066371** | **DHKHMT16B** |
| **Nguyễn Văn Hiền** | **20116451** | **DHKHMT16A** |

**MINISTRY OF INDUSTRY AND TRADE INDUSTRIAL**

**UNIVERSITY OF HO CHI MINH CITY FACULTY OF INFORMATION TECHNOLOGY**

-----🙞🙜🕮🙞🙜-----



**COMPUTER VISION AND APPLICATIONS**

**TEXTS AS IMAGES IN PROMPT TUNING FOR MULTI-LABEL IMAGE RECOGNITION**

**Supervisor: Dr. Nguyen Chi Kien**

**Student Name Student code Class**

|  |  |  |
| --- | --- | --- |
| **Đoàn Ánh Dương** | **20082631** | **DHKHMT16B** |
| **Nguyễn Hữu Hải** | **20053571** | **DHKHMT16A** |
| **Nguyễn Thị Hồng Vân** | **20003445** | **DHKHMT16A** |
| **Trần Thanh Đức** | **20066371** | **DHKHMT16B** |
| **Nguyễn Văn Hiền** | **20116451** | **DHKHMT16A** |

# ABSTRACT

Prompt tuning has been utilized as an effective method to adapt large pre-trained models in vision and language (e.g., CLIP) for specific tasks in situations with limited data or labels. In this project, we employ Prompt tuning techniques to process text as images in multi-label image recognition and introduce the TaI (Texts as Images) prompting method. Unlike image data, textual descriptions are easy to collect and their class labels can be directly identified. Specifically, we apply the TaI prompting method to multi-label image recognition, where real-world descriptive sentences are used instead of images for prompt tuning. Furthermore, with TaI, we introduce a double-grained prompt tuning method (TaI-DPT) to extract both coarse-grained and fine-grained embeddings, aiming to enhance multi-label recognition performance.

**Keywords:** Texts as Images, prompt tuning, double-grained prompt tuning, CLIP, multi-label image recognition

# LỜI CAM ĐOAN

Chúng em xin cam đoan việc nghiên cứu đề tài được thực hiện cùng với giáo viên hướng dẫn khóa luận và chưa được sử dụng để báo cáo cho bất kì đồ án hay môn học nào.

Chúng em xin cam đoan mọi sự giúp đỡ đã được cảm ơn, các thông tin trích dẫn đã được ghi chú đầy đủ và bảo đảm rõ nguồn gốc không xâm phạm quyền tác giả.

TP. Hồ Chí Minh, tháng 5 năm 2024

# LỜI CẢM ƠN

Bài tiểu luận về đề tài “Texts as Images in Prompt Tuning for Multi-Label Image Recognition” thuộc bộ môn thị giác máy tính và ứng dụng là kết quả của quá trình học tập, tiếp thu kiến thức tại trường, lớp và cả những tìm tòi, nghiên cứu riêng của nhóm em và sự chỉ dạy tận tình của thầy Nguyễn Chí Kiên - người đã trực tiếp hướng dẫn nhóm em trong môn học này. Do vậy, qua đây nhóm em xin phép được gửi lời cảm ơn chân thành nhất tới quý thầy.

Mặc dù đã dành nhiều thời gian và nỗ lực để hoàn thành bài tiểu luận này, nhưng do sự hạn chế về mặt kiến thức nên bài làm khó tránh khỏi những thiếu sót. Nhóm em kính mong nhận được những lời góp ý của quý thầy, cô để bài làm ngày càng hoàn thiện hơn.

Chúng em xin chân thành cảm ơn!

TP. Hồ Chí Minh, tháng 5 năm 2024

# NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

TP. Hồ Chí Minh, tháng 5 năm 2024

GIẢNG VIÊN HƯỚNG DẪN

TS. NGUYỄN CHÍ KIÊN

# NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

......................................................................................................................................

TP. Hồ Chí Minh, tháng 5 năm 2024.

**GIẢNG VIÊN PHẢN BIỆN**

# MỤC LỤC

[ABSTRACT 3](#_Toc167488024)

[LỜI CAM ĐOAN 1](#_Toc167488025)

[LỜI CẢM ƠN 2](#_Toc167488026)

[NHẬN XÉT CỦA GIÁO VIÊN HƯỚNG DẪN 3](#_Toc167488027)

[NHẬN XÉT CỦA GIÁO VIÊN PHẢN BIỆN 4](#_Toc167488028)

[MỤC LỤC 5](#_Toc167488029)

[DANH MỤC BẢNG BIỂU 8](#_Toc167488030)

[MỤC LỤC HÌNH ẢNH 9](#_Toc167488031)

[DANH MỤC TỪ VIẾT TẮT 10](#_Toc167488032)

[CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI 12](#_Toc167488033)

[1.1 Lý do chọn đề tài 12](#_Toc167488034)

[1.2 Mục tiêu và nhiệm vụ nghiên cứu 13](#_Toc167488035)

[Bài toán đặt ra 13](#_Toc167488036)

[Yêu cầu phi chức năng 13](#_Toc167488037)

[1.3 Đối tượng nghiên cứu 14](#_Toc167488038)

[Dữ liệu hình ảnh 14](#_Toc167488039)

[Mô hình huấn luyện 14](#_Toc167488040)

[Kỹ thuật Prompt tuning 14](#_Toc167488041)

[Dữ liệu văn bản 14](#_Toc167488042)

[Đánh giá hiệu suất 14](#_Toc167488043)

[1.4 Phạm vi nghiên cứu 15](#_Toc167488044)

[1.5 Phương pháp nghiên cứu 15](#_Toc167488045)

[1.6. Ý nghĩa khoa học và thực tiễn 15](#_Toc167488046)

[Về mặt khoa học 15](#_Toc167488047)

[Về mặt thực tiễn 16](#_Toc167488048)

[CHƯƠNG 2: CƠ SỞ LÝ THUYẾT 17](#_Toc167488049)

[2.1 Giới thiệu Prompt tuning 17](#_Toc167488050)

[2.2 Multi-Label Image Recognition 17](#_Toc167488051)

[2.2.1 Multi-Label Image Recognition (nhận diện hình ảnh đa nhãn) 17](#_Toc167488052)

[2.2.2 Multi-Label Recognition from Few-shot Samples (nhận diện đa nhãn từ mẫu Few-shot) 18](#_Toc167488053)

[2.2.3 Multi-Label Recognition from Partial-label Data (nhận diện đa nhãn từ dữ liệu Partial-label) 18](#_Toc167488054)

[2.3 TaI prompting 18](#_Toc167488055)

[2.4 Text -as-Image trong quy trình tinh chỉnh gợi ý hai mức độ 19](#_Toc167488056)

[2.5 Hàm mất mát được sử dụng để huấn luyện mô hình 21](#_Toc167488057)

[CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM 23](#_Toc167488058)

[3.1 Giới thiệu dữ liệu 23](#_Toc167488059)

[3.2 Tiến hành huấn luyện 23](#_Toc167488060)

[3.3 Kết quả huấn luyện 23](#_Toc167488061)

[CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN 25](#_Toc167488062)

[4.1 Kết quả đề tài 25](#_Toc167488063)

[CHƯƠNG 5: TÀI LIỆU THAM KHẢO 26](#_Toc167488064)

# DANH MỤC BẢNG BIỂU

[Bảng 1: Siêu tham số của mô hình 24](#_Toc167488355)

[Bảng 2: Kết quả huấn luyện phương pháp TaI-DPT 24](#_Toc167488356)

# MỤC LỤC HÌNH ẢNH

[Hình 1: Quy trình đào tạo và thử nghiệm chuyển đổi văn bản thành hình ảnh 19](#_Toc167488607)

# DANH MỤC TỪ VIẾT TẮT

|  |  |  |
| --- | --- | --- |
| **Từ viết tắt** | **Từ đầy đủ** | **Nghĩa** |
| AI | Artificial Intelligence | Trí tuệ nhân tạo |
| NLP | **Natural Language Processing** | Xử lý ngôn ngữ tự nhiên |
| MS COCO | Microsoft Common Objects in Context | Tập dữ liệu lớn được phát triển bởi Microsoft |
| Pascal VOC | **Pattern Analysis, Statistical Modelling, and Computational Learning Visual Object Classes** | Tập dữ liệu chứa hình ảnh với các đối tượng được chú thích |
| CNN | Convolutional Neural Network | Mạng Nơ-ron Tích Chập |
| VOC2007 | Visual Object Classes Challenge 2007 | Tập hợp các hình ảnh đã được chú thích |
| NUS-WIDE | National University of Singapore-Wide | Tập dữ liệu được gắn nhãn với một số lượng lớn các từ khóa (tags) mô tả nội dung của hình ảnh |
| TaI-DPT | Task-aware Information Decoupling Pre-trained Transformer | Phương pháp để cải thiện hiệu suất của các mô hình Transformer đã được huấn luyện trước |
| CLIP | Contrastive Language-Image Pre-training | Mô hình học sâu biểu diễn văn bản - hình ảnh |
| CHAMP | Confidence-Weighted Hierarchical Aggregated Multi-Prototype Network | Kỹ thuật phân loại dựa trên mạng nơ-ron đa tầng |
| LaSO | Label Set Operations | Các phép toán trên tập hợp nhãn |
| VL | Vision-Language | Tương tác giữa Thị giác và Ngôn ngữ |
| CoOp | Contrastive Optimized Prototyping | Tối ưu hóa Prototype theo Cách Tương Phản |
| ResNet-50 | Residual Network | Mạng còn lại bao gồm 50 lớp |
| SGD | Stochastic Gradient Descent | Tối ưu hóa Gradient Ngẫu nhiên |
| Mean AP | Mean Average Precision | Độ chính xác trung bình |

# CHƯƠNG 1: TỔNG QUAN ĐỀ TÀI

## Lý do chọn đề tài

Công nghệ càng hiện đại, con người càng phát triển. Nhu cầu học hỏi, muốn tìm tòi khám phá thế giới của con người càng lớn. Để đáp ứng được vấn đề này, phương pháp “Texts as Images in Prompt Tuning for Multi-Label Image Recognition” đã ra đời. Trước hết, nghiên cứu này tận dụng mối quan hệ ngày càng mật thiết giữa văn bản và hình ảnh trong các mô hình AI hiện đại, mở ra nhiều tiềm năng ứng dụng thực tiễn. Sử dụng văn bản như các thông điệp điều chỉnh (prompt tuning) có thể giúp cải thiện đáng kể hiệu suất của các mô hình nhận diện hình ảnh, đặc biệt là trong bối cảnh đa nhãn, từ đó tăng độ chính xác và độ tin cậy của các hệ thống AI khi nhận dạng nhiều đối tượng hoặc khái niệm trong một hình ảnh duy nhất. Kỹ thuật này không chỉ mang tính đa dụng và linh hoạt, mà còn có thể áp dụng cho nhiều loại dữ liệu và ngữ cảnh khác nhau.

Hơn nữa, đề tài này còn có ý nghĩa đóng góp đáng kể vào sự phát triển của các kỹ thuật AI hiện đại, mở ra những hướng đi mới trong việc kết hợp giữa xử lý ngôn ngữ tự nhiên (NLP) và thị giác máy tính (computer vision). Công nghệ này cũng có ứng dụng thực tiễn cao trong nhiều lĩnh vực khác nhau như y tế, an ninh, thương mại điện tử, và giáo dục. Ví dụ, trong y tế, việc nhận diện nhiều loại bệnh hoặc triệu chứng từ hình ảnh y học có thể được cải thiện đáng kể; trong thương mại điện tử, nhận diện nhiều loại sản phẩm hoặc thuộc tính sản phẩm từ hình ảnh có thể nâng cao trải nghiệm người dùng và hiệu quả kinh doanh.

Ngoài ra, đề tài này thúc đẩy nghiên cứu liên ngành, yêu cầu sự kết hợp giữa học sâu, thị giác máy tính, xử lý ngôn ngữ tự nhiên, và học chuyển giao (transfer learning), qua đó khuyến khích sự hợp tác giữa các nhà nghiên cứu trong các lĩnh vực khác nhau. Cuối cùng, việc sử dụng văn bản như các hình ảnh trong việc điều chỉnh prompt cho nhận diện hình ảnh đa nhãn là một hướng tiếp cận mới mẻ và đầy thách thức, đòi hỏi sự sáng tạo trong cách tiếp cận và các kỹ thuật tiên tiến, kiến thức sâu rộng về cả hai lĩnh vực NLP và thị giác máy tính.

Xét thực tế thì nghiên cứu này có thể được ứng dụng rộng rãi trong các lĩnh vực khác nhau của đời sống công nghệ, mang lại giá trị cao cho người sử dụng. Vì vậy nhóm em quyết định chọn đề tài “**Texts as Images in Prompt Tuning for Multi-Label Image Recognition**” làm hướng nghiên cứu cho bài báo cáo tiểu luận của nhóm.

## Mục tiêu và nhiệm vụ nghiên cứu

Mục tiêu của nghiên cứu này là phát triển một phương pháp mới sử dụng văn bản như các hình ảnh trong việc điều chỉnh prompt (prompt tuning) để cải thiện hiệu suất của các mô hình nhận diện hình ảnh đa nhãn (multi-label image recognition).

## Bài toán đặt ra

Bài toán chính của nghiên cứu này là làm thế nào để sử dụng văn bản như các hình ảnh trong việc điều chỉnh prompt nhằm cải thiện nhận diện hình ảnh đa nhãn. Điều này bao gồm việc tạo ra một mô hình có khả năng:

* **Xử lý đồng thời thông tin từ văn bản và hình ảnh:** Tận dụng thông tin ngữ cảnh từ văn bản để nâng cao khả năng nhận diện và phân loại nhiều nhãn trong một hình ảnh.
* **Tương thích với các bộ dữ liệu lớn và đa dạng:** Đảm bảo mô hình hoạt động tốt trên các bộ dữ liệu lớn và có sự đa dạng về nội dung, bối cảnh, và ngữ cảnh.
* **Tích hợp và tối ưu hóa các kỹ thuật học sâu:** Sử dụng các kỹ thuật học sâu tiên tiến như học chuyển giao (transfer learning) và học tăng cường (reinforcement learning) để tối ưu hóa mô hình.

## Yêu cầu phi chức năng

* **Hiệu suất tính toán:** Mô hình cần được tối ưu hóa để có thể chạy trên các hệ thống tính toán phổ biến với thời gian xử lý hợp lý.
* **Khả năng mở rộng:** Mô hình phải có khả năng mở rộng để xử lý các bộ dữ liệu lớn hơn và phức tạp hơn trong tương lai.
* **Tính dễ hiểu và giải thích:** Kết quả của mô hình cần dễ hiểu và có thể giải thích được, đặc biệt là trong các ứng dụng thực tế như y tế và an ninh.
* **Tính khả dụng và tái sử dụng:** Mô hình và các phương pháp phát triển cần được thiết kế sao cho có thể tái sử dụng và áp dụng cho các bài toán khác mà không cần phải thay đổi nhiều.
* **Bảo mật và quyền riêng tư:** Đảm bảo rằng mô hình tuân thủ các tiêu chuẩn bảo mật và quyền riêng tư, đặc biệt là khi xử lý dữ liệu nhạy cảm như hình ảnh y tế hoặc thông tin cá nhân.

## Đối tượng nghiên cứu

## Dữ liệu hình ảnh

Trong đề tài này chúng em sử dụng các bộ dữ liệu hình ảnh chứa nhiều đối tượng hoặc nhãn trong mỗi hình ảnh, như MS COCO, Pascal VOC, và Open Images. Những bộ dữ liệu này cung cấp nền tảng để huấn luyện và kiểm tra hiệu suất của mô hình nhận diện hình ảnh đa nhãn.

## Mô hình huấn luyện

Các mô hình học sâu tiên tiến như CNN (Convolutional Neural Networks), Transformers, và các mô hình kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên. Các mô hình này sẽ được điều chỉnh và tối ưu hóa để sử dụng văn bản làm prompt trong quá trình nhận diện hình ảnh.

## Kỹ thuật Prompt tuning

Phương pháp điều chỉnh prompt, bao gồm cách thức chuyển đổi văn bản thành các hình ảnh hoặc biểu diễn ngữ cảnh để cải thiện hiệu suất của mô hình. Điều này bao gồm nghiên cứu các kỹ thuật mã hóa văn bản và tích hợp chúng vào mô hình thị giác máy tính.

## Dữ liệu văn bản

Các nguồn dữ liệu văn bản liên quan đến hình ảnh, như chú thích hình ảnh, mô tả văn bản, và các văn bản ngữ cảnh khác. Các nguồn này cung cấp thông tin bổ sung để hỗ trợ quá trình điều chỉnh prompt và nâng cao khả năng nhận diện của mô hình.

## Đánh giá hiệu suất

Các tiêu chí và phương pháp đánh giá hiệu suất của mô hình, bao gồm độ chính xác (precision), độ nhạy (recall), F1-score, và các chỉ số khác để đánh giá khả năng nhận diện đa nhãn của mô hình trong các bối cảnh và bộ dữ liệu khác nhau.

## Phạm vi nghiên cứu

Đề tài "Texts as Images in Prompt Tuning for Multi-Label Image Recognition" (TaI-DPT) sử dụng các tập dữ liệu phổ biến như VOC2007, MS-COCO và NUS-WIDE để tối ưu hóa việc nhận dạng hình ảnh đa nhãn. Nó đề xuất một phương pháp mới nhằm cải thiện hiệu suất so với các phương pháp zero-shot truyền thống trên các bộ dữ liệu phổ biến này.

## Phương pháp nghiên cứu

Phương pháp TaI-DPT kết hợp dữ liệu chú thích văn bản từ các bộ dữ liệu như MS-COCO và OpenImages, sau đó sử dụng chúng để điều chỉnh mô hình dựa trên kiến trúc CLIP. Phương pháp này bao gồm ba giai đoạn: chuyển đổi văn bản thành hình ảnh sử dụng các chú thích văn bản từ các bộ dữ liệu và chuyển đổi chúng thành dạng hình ảnh, điều chỉnh mô hình sử dụng dữ liệu văn bản đã chuyển đổi để điều chỉnh mô hình dựa trên kiến trúc CLIP, huấn luyện và đánh giá mô hình trên các tập dữ liệu đa nhãn. Kết quả được so sánh với các phương pháp hiện tại để đánh giá hiệu suất.

## 1.6. Ý nghĩa khoa học và thực tiễn

## Về mặt khoa học

* **Mở rộng kiến thức về nhận diện hình ảnh đa nhãn:** Nghiên cứu này cung cấp cái nhìn mới về cách sử dụng văn bản như các hình ảnh trong quá trình điều chỉnh prompt để cải thiện hiệu suất của mô hình nhận diện hình ảnh đa nhãn. Điều này mở ra cánh cửa cho việc phát triển các phương pháp mới và tiên tiến trong lĩnh vực này.
* **Kết hợp giữa thị giác máy tính và xử lý ngôn ngữ tự nhiên:** Phương pháp được đề xuất trong nghiên cứu này đề xuất cách kết hợp giữa hai lĩnh vực trí tuệ nhân tạo: thị giác máy tính và xử lý ngôn ngữ tự nhiên. Điều này cung cấp cơ sở cho việc phát triển các giải pháp đa dạng và tích hợp sâu sắc giữa hai lĩnh vực này trong tương lai.
* **Nâng cao hiệu suất và tổng quát hóa của mô hình:** Nghiên cứu này nhằm tối ưu hóa hiệu suất của mô hình nhận diện hình ảnh đa nhãn thông qua việc sử dụng văn bản làm prompt. Kết quả có thể đóng góp vào việc nâng cao hiệu suất và tổng quát hóa của các mô hình tương tự trong tương lai.

## Về mặt thực tiễn

* **Ứng dụng trong y tế:** Phương pháp nhận diện hình ảnh đa nhãn cải thiện có thể được áp dụng trong việc nhận diện và phân loại triệu chứng từ hình ảnh y học, hỗ trợ cho việc chuẩn đoán và điều trị bệnh.
* **Ứng dụng trong an ninh:** Công nghệ nhận diện hình ảnh đa nhãn có thể được sử dụng để phát hiện đối tượng và hành vi nghi ngờ trong các hệ thống giám sát, cải thiện an ninh và giám sát công cộng.
* **Ứng dụng trong thương mại điện tử:** Khả năng nhận diện sản phẩm và thuộc tính từ hình ảnh giúp cải thiện trải nghiệm mua sắm trực tuyến và tăng cường khả năng tìm kiếm và tư vấn sản phẩm cho người tiêu dùng.
* **Ứng dụng trong giáo dục:** Công nghệ nhận diện hình ảnh đa nhãn có thể được áp dụng trong việc phân loại và nhận diện nội dung học tập, hỗ trợ cho việc tự động hóa quá trình đánh giá và phản hồi trong giáo dục trực tuyến và từ xa.

# CHƯƠNG 2: CƠ SỞ LÝ THUYẾT

## Giới thiệu Prompt tuning

Trong những năm gần đây, có những tiến bộ nhanh chóng trong các mô hình thị giác-ngôn ngữ (vision-language) lớn. Các mô hình được huấn luyện trước cũng như hiệu suất đáng kể của chúng trong các nhiệm vụ thị giác ứng dụng (bao gồm nhận diện đối tượng, phân loại hình ảnh, phát hiện vật thể, phân đoạn hình ảnh, và nhiều ứng dụng khác). Một mô hình thị giác - ngôn ngữ được huấn luyện trước thường bao gồm các bộ mã hóa dữ liệu, và việc sử dụng hàm mất mát đối chiếu hình ảnh - văn bản để căn chỉnh các embedding của hình ảnh và văn bản vào một không gian chung đang trở nên ngày càng phổ biến. Khi áp dụng vào các nhiệm vụ ứng dụng trong các tình huống hạn chế dữ liệu hoặc nhãn, việc tinh chỉnh toàn bộ mô hình thường không hiệu quả do độ phức tạp cao của nó. Do đó, prompt tuning ra đời như một mô hình học tập tiết kiệm tham số tiêu biểu để điều chỉnh các mô hình thị giác - ngôn ngữ lớn cho phù hợp với các nhiệm vụ thị giác ứng dụng.

## Multi-Label Image Recognition

## Multi-Label Image Recognition (nhận diện hình ảnh đa nhãn)

Nhận diện hình ảnh đa nhãn nhằm mục đích nhận diện tất cả các danh mục đối tượng hoặc các khái niệm trong một hình ảnh đầu vào. Các mô-đun khác nhau đã được giới thiệu để đại diện tốt hơn cho các mối quan hệ giữa các lớp, và các hàm mất mát phân loại hiện đại đã được sử dụng để làm cho việc học mô hình dễ dàng hơn. Để mô hình hóa sự phụ thuộc của nhãn, CNN-RNN giới thiệu các mạng nơ-ron hồi quy, ví dụ như RNN và LSTM, để dự đoán các lớp xuất hiện theo cách tuần tự. Sử dụng các mô-đun tích chập đồ thị để học mối quan hệ giữa các nhãn lớp. Phương pháp CHAMP (Confidence-Weighted Hierarchical Aggregated Multi-Prototype Network) được sử dụng để đo lường mức độ nghiêm trọng của việc phân loại sai lầm bằng cách xây dựng một cây phân cấp cụ thể cho từng lĩnh vực dựa trên mối quan hệ giữa các danh mục, trong đó mỗi lớp được liên kết với một nút cây, nhằm cải thiện tính ổn định của mô hình. Mặc dù hiệu quả, nhưng những phương pháp này yêu cầu một số lượng đáng kể hình ảnh được gán nhãn để học đủ các mối quan hệ giữa các danh mục. Trong các trường hợp giới hạn dữ liệu hoặc nhãn, ví dụ như dữ liệu ít hoặc dữ liệu nhãn một phần, việc này sẽ khó khăn cho các mô hình này học tốt như mong đợi. Các hàm mất mát được thiết kế đặc biệt cũng gặp khó khăn trong việc đạt được các cải tiến đáng kể khi học với dữ liệu hạn chế.

## Multi-Label Recognition from Few-shot Samples (nhận diện đa nhãn từ mẫu Few-shot)

Để khai thác tốt hơn số lượng mẫu ít, LaSO tổng hợp các mẫu bằng cách thao tác các đặc trưng của các hình ảnh huấn luyện được ghép cặp. Các cách thao tác bộ nhãn khác nhau được sử dụng để huấn luyện mô hình, dẫn đến các đặc trưng phân biệt có tính tổng quát. giới thiệu một khung meta-learning để học tốt hơn các nhiệm vụ trước đó và tổng quát hóa đến các nhiệm vụ mới, và tận dụng số lượng nhãn như một thông tin hữu ích cho việc học.

## Multi-Label Recognition from Partial-label Data (nhận diện đa nhãn từ dữ liệu Partial-label)

Partial-label đề cập đến các trường hợp mà một số nhãn không được xác định. Mặc dù đã đạt được tiến bộ đáng kể, nhưng vẫn là một vấn đề thách thức để học nhận diện hình ảnh đa nhãn trong các chế độ hạn chế hình ảnh hoặc nhãn. Dựa trên các mô hình VL (thị giác – ngôn ngữ) đã được huấn luyện trước, bài báo này đề xuất học các prompt từ văn bản thay vì hình ảnh, qua đó cung cấp một góc nhìn mới nhưng bổ sung cho việc xử lý nhận diện hình ảnh đa nhãn trong điều kiện tài nguyên thấp.

## TaI prompting

Ở đây, nhóm giới thiệu phương pháp gợi ý văn bản dưới dạng hình ảnh (Text-as-Image, TaI) để điều chỉnh các mô hình thị giác-ngôn ngữ (VL) đã được huấn luyện trước cho nhiệm vụ nhận dạng hình ảnh đa nhãn. Phương pháp gợi ý TaI của chúng em chỉ sử dụng các văn bản tự do dễ truy cập làm dữ liệu huấn luyện để học cách gợi ý hiệu quả cho các tác vụ nhận dạng đa nhãn sau này.

A close-up of a bus

Description automatically generated

Hình 1: Quy trình đào tạo và thử nghiệm chuyển đổi văn bản thành hình ảnh

Hình 1 minh họa thiết kế của khung TaI-DPT mà chúng em đề xuất, bao gồm các giai đoạn huấn luyện và kiểm tra. Trong giai đoạn huấn luyện, chúng em học các gợi ý (prompts) chỉ với sự giám sát từ các văn bản. Hai bản sao giống hệt nhau của bộ mã hóa văn bản EncT từ CLIP đã được tiền huấn luyện được sử dụng để mã hóa các gợi ý và dữ liệu văn bản tương ứng. Chúng em giới thiệu hai loại gợi ý có thể huấn luyện (tức là gợi ý toàn cục và gợi ý địa phương) để thu được các nhúng lớp (class embeddings) toàn cục và khu vực.

Chiến lược lọc danh từ được sử dụng để tạo ra các nhãn giả cho mỗi mô tả văn bản, điều này được áp dụng để giám sát các điểm phân loại thu được bằng cách tính toán độ tương đồng cosine giữa các nhúng lớp và các đặc trưng văn bản. Chỉ các tham số trong các gợi ý được tối ưu hóa trong giai đoạn huấn luyện, trong khi cả hai bộ mã hóa văn bản đều được giữ cố định. Trong giai đoạn kiểm tra, các nhúng lớp được thu thập bằng cách mã hóa hai bộ gợi ý đã học với bộ mã hóa văn bản EncT như trong quá trình huấn luyện, trong khi nguồn đầu vào khác chuyển từ mô tả văn bản sang hình ảnh kiểm tra. Bộ mã hóa hình ảnh EncI đã được tiền huấn luyện từ CLIP được sử dụng để trích xuất các đặc trưng toàn cục và dày đặc của mỗi hình ảnh kiểm tra, sau đó độ tương đồng cosine được tính giữa các đặc trưng và các nhúng lớp của các gợi ý toàn cục và địa phương. Kết quả phân loại cuối cùng được thu bằng cách kết hợp các điểm cosine toàn cục và địa phương.

## Text -as-Image trong quy trình tinh chỉnh gợi ý hai mức độ

Một prompt được định nghĩa như sau:

(1)

trong đó: là chỉ số lớp, biểu thị nhúng từ của tên lớp thứ . Đối với j{1;...;M}, là một nhúng từ có thể học có kích thước giống với kích thước của các nhúng từ bình thường trong từ vựng. Giống như các phương pháp trước đây, ví dụ như CoOp, các prompts được học bằng cách tối đa hóa xác suất phân loại mỗi hình ảnh vào đúng lớp của nó:

(2)

trong đó biểu thị hình ảnh huấn luyện có nhãn và ⟨⋅,⋅⟩ tính toán độ tương đồng cosine.

Để nắm bắt khả năng phân biệt của các đặc trưng toàn cục cũng như học từ các đặc trưng chi tiết, nhóm giới thiệu phương pháp điều chỉnh prompt kép (tức là TaI-DPT) sử dụng hai bộ prompt để xử lý các đặc trưng toàn cục (tức là mức độ thô nhất) và địa phương (tức là mức độ chi tiết) tương ứng trong hai nhánh song song. Prompt cấp kép được định nghĩa như sau:

(3)

trong đó ​ và là các nhúng có thể học được kết hợp với nhúng từ ​ của lớp để thu được prompt toàn cục ​ và prompt địa phương ​ ​ tương ứng. Các chuỗi trong phương trình (3) được đưa vào một bản sao của bộ mã hóa văn bản EncT của CLIP để tạo ra các nhúng lớp toàn cục và địa phương cho mỗi lớp, tức là và , và được khuyến khích để tương quan với các đặc trưng toàn cục và địa phương tương ứng.

Để giữ lại các đặc trưng vùng chi tiết cho hình ảnh đầu vào, chúng em duy trì bản đồ đặc trưng trước lớp gộp chú ý của CLIP. Đối với mô tả văn bản đầu vào,nhóm giữ lại các đặc trưng tuần tự của toàn bộ câu thay vì chỉ các đặc trưng token <EOS>.

trong đó biểu thị một mô tả văn bản huấn luyện. là các đặc trưng hình ảnh và văn bản toàn cục được trích xuất. và là các đặc trưng hình ảnh dày đặc được làm phẳng và các đặc trưng tuần tự token tương ứng, trong đó biểu thị kích thước không gian phẳng của đặc trưng thị giác và biểu thị độ dài của các token văn bản. Độ tương đồng toàn cục và địa phương được tính bằng:

trong đó biểu thị đặc trưng ngôn ngữ trong quá trình huấn luyện hoặc đặc trưng thị giác trong quá trình kiểm tra, và biểu thị hoặc tương ứng. Thông tin trong nhánh địa phương được tổng hợp theo cách trọng số không gian:

trong đó *τs* điều chỉnh mức độ tập trung vào một vị trí cụ thể. ​ và ​ được tối ưu hóa bởi các thuật ngữ mất mát ​ và ​ tương ứng, mà chúng em sẽ thảo luận trong phần tiếp theo. Và trong giai đoạn kiểm tra, và được kết hợp để thu được điểm phân loại cuối cùng.

## Hàm mất mát được sử dụng để huấn luyện mô hình

Mục tiêu học tập tổng thể được định nghĩa là ​trong đó và là các thuật ngữ mất mát cho nhúng văn bản toàn cục và các token văn bản địa phương tương ứng. Chúng em áp dụng mất mát xếp hạng để đo lường sự khác biệt giữa điểm phân loại và các nhãn giả trên cơ sở sự thật. Cụ thể, và ​ được công thức hóa như sau:

trong đó và là các độ tương đồng toàn cục và tổng hợp cục bộ được mô tả trong mục 3.3, là biên kiểm soát mức độ tương đồng của các lớp tích cực cao hơn bao nhiêu so với các lớp tiêu cực. Trong quá trình huấn luyện, chúng em tối thiểu hóa mục tiêu tổng thể với các bộ mã hóa văn bản cố định, bằng cách tối ưu hóa các prompts toàn cục và địa phương.

Mất mát entropy chéo nhị phân và các biến thể của nó như mất mát bất đối xứng đã cho thấy kết quả đáng chú ý trong các nhiệm vụ phân loại. Chúng thường đi kèm với hàm sigmoid σ(x) =​ để chuyển đổi đầu ra của mô hình thành xác suất. Tuy nhiên, chúng em nhận thấy rằng việc tối ưu hóa trực tiếp xác suất σ(p) dẫn đến khoảng cách hiệu suất giữa các văn bản huấn luyện và hình ảnh kiểm tra. Chúng em cho rằng hiện tượng này là do khoảng cách giữa các phương thức thị giác và ngôn ngữ. Vì vậy, chúng em coi mất mát xếp hạng là một cách giám sát linh hoạt và phù hợp hơn trong bối cảnh có khoảng cách phương thức.

# CHƯƠNG 3: KẾT QUẢ THỰC NGHIỆM

## Giới thiệu dữ liệu

Để đánh giá phương pháp TaI-DPT của chúng em, chúng em tiến hành các thí nghiệm trên VOC2007, MS-COCO, và NUS-WIDE. VOC2007 chứa 20 danh mục phổ biến, và chúng em tạo ra tập huấn luyện / thử nghiệm dựa trên sự chia tách chính thức của train val / test (5,011 hình ảnh / 4,952 hình ảnh). MS-COCO bao gồm 80 danh mục với 82,081 hình ảnh huấn luyện và 40,504 hình ảnh kiểm tra. NUS-WIDE bao gồm 81 danh mục với 161,789 hình ảnh cho việc huấn luyện. Chúng em sử dụng tập kiểm tra của nó (107,859 hình ảnh) để đánh giá phương pháp của chúng em . Ngoài ra, đối với VOC2007 và MS-COCO, các nguồn dữ liệu ngôn ngữ là các chú thích từ MS-COCO. Đối với NUS-WIDE, chúng em giới thiệu các mô tả cục bộ từ OpenImages, có phạm vi nội dung rộng hơn, để bao quát tất cả các khái niệm trong NUS-WIDE.

## Tiến hành huấn luyện

Nhóm sử dụng CLIP ResNet-50 làm bộ mã hóa trực quan và sử dụng CLIP Transformer làm bộ mã hóa văn bản. Trong quá trình đào tạo, các thông số của hai bộ mã hóa được giữ nguyên và learnable prompts có thể học được mới được tối ưu hóa.

## Kết quả huấn luyện

Trong quá trình huấn luyện nhóm áp dụng bộ tối ưu hóa SGD, và số epoch huấn luyện được đặt là 20 cho tất cả các bộ dữ liệu. Tỷ lệ học cho MS-COCO, VOC2007, và NUS-WIDE được khởi tạo theo cách thực nghiệm với 1e-4, 1e-4, và 1e-3, và giảm theo quy tắc cosine annealing. Đối với mất mát, chúng em chọn m = 1, và co giãn các tham số p và p’ bằng một hệ số 4. τs được đặt là 0.02 thông qua việc xác nhận.

* Thiết kế siêu tham số:

|  |  |
| --- | --- |
| Epoch | 20 |
| Learning rate | 0.0001 |
| Learning rate\_ scheduler | cosine |
| Gamma | 0.1 |
| Weight decay | 0.0005 |
| Momemtum | 0.9 |

Bảng 1: Siêu tham số của mô hình

* Kết quả sau khi huấn luyện:

|  |  |  |
| --- | --- | --- |
| **Datasets** | **Method** | **Mean AP** |
| PascalVOC 2007 | TaI-DPT | 88.3 |
| MS-COCO | TaI-DPT | 65.1 |
| NUSWIDE | TaI-DPT | 46.5 |

Bảng 2: Kết quả huấn luyện phương pháp TaI-DPT

Nhận xét: Phương pháp TaI-DPT đạt Mean AP cao là 88.3 trên PascalVOC 2007. Điều này cho thấy mô hình hoạt động rất tốt trên bộ dữ liệu này, có thể do các danh mục đối tượng được định nghĩa rõ ràng và số lượng hạn chế. Mean AP giảm xuống.

còn 65.1 trên bộ dữ liệu MS-COCO do MS-COCO có nhiều danh mục đối tượng hơn và các cảnh phức tạp hơn với nhiều đối tượng trong mỗi hình ảnh. Trên bộ dữ liệu NUSWIDE, Mean AP giảm xuống còn 46.5. Đây là mức giảm đáng kể so với các điểm số trên PascalVOC và MS-COCO. NUSWIDE có nhiều danh mục rất đa dạng và các khái niệm không được định nghĩa chặt chẽ, làm cho nhiệm vụ trở nên khó khăn hơn. Thêm vào đó, các hình ảnh web có thể có chất lượng khác nhau và nhiều nhiễu.

# CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

## Kết quả đề tài

Trong bài báo cáo này, nhóm đề xuất một quan điểm mới về việc coi văn bản là hình ảnh trong điều chỉnh nhanh chóng (tức là TaI), học lời nhắc từ các tính năng phân biệt đối xử của mô tả văn bản. So với các phương pháp điều chỉnh nhanh trước đây được đào tạo bằng hình ảnh, TaI của chúng tôi được hưởng lợi từ khả năng truy cập dễ dàng của các văn bản giàu nội dung có thể mở rộng, cho phép điều chỉnh nhanh chóng cho các tác vụ thị giác (ví dụ*:* nhận dạng hình ảnh đa nhãn)

Hướng phát triển tiếp theo nhóm tiếp tục khám phá cách tạo mô tả văn bản tinh vi hơn và tối ưu hóa sự tích hợp với các phương pháp điều chỉnh prompt dựa trên hình ảnh, bao gồm cải thiện việc thu thập dữ liệu, nghiên cứu và triển khai các phương pháp tiên tiến hơn để giải quyết vấn đề của dữ liệu đa dạng và tài nguyên tính toán.

# CHƯƠNG 5: TÀI LIỆU THAM KHẢO

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. arXiv preprint arXiv:2204.14198, 2022.

[2] Amit Alfassy, Leonid Karlinsky, Amit Aides, Joseph Shtok, Sivan Harary, Rogerio Feris, Raja Giryes, and Alex M Bronstein. Laso: Label-set operations networks for multi-label few-shot learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6548–6557, 2019

[3] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2021.

[4] Steven Bird, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”, 2009

[5] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. In Proceedings of the AAAI conference on artificial intelligence, volume 36, pages 339–346, 2022.

[6] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In Proceedings of the IEEE/CVF international conference on computer vision, pages 522–531, 2019

[7] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In ICME 2019

[8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In CVPR, 2019

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In Proceedings of the ACM international conference on image and video retrieval, pages 1–9, 2009

[10] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 647–657, 2019

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010.

[12] Bin-Bin Gao and Hong-Yu Zhou. Multi-label image recognition with multi-class attentional regions. arXiv preprint arXiv:2007.01755, 2020

[13] Chunjiang Ge, Rui Huang, Mixue Xie, Zihang Lai, Shiji Song, Shuang Li, and Gao Huang. Domain adaptation via prompt learning. arXiv preprint arXiv:2202.06687, 2022

[14] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for mul tilabel image annotation. arXiv preprint arXiv:1312.4894, 2013.

[15] Shiyi He, Chang Xu, Tianyu Guo, Chao Xu, and Dacheng Tao. Reinforced multi-label image classification by exploring curriculum. In AAAI, 2018

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916. PMLR, 2021.

[17] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. arXiv preprint arXiv:2203.12119, 2022

[18] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html, 2017.

[19] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208, 2021

[20] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence ´ Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740–755. Springer, 2014

[21] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. arXiv preprint arXiv:2205.14865, 2022.

[22] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337–2348, 2022.

[23] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16816–16825, 2022.

[24] Yuanhan Zhang, Kaiyang Zhou, and Ziwei Liu. Neural prompt search. arXiv preprint arXiv:2206.04673, 2022

[25] Renrui Zhang, Zhang Wei, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tipadapter: Training-free adaption of clip for few-shot classification. arXiv preprint arXiv:2207.09519, 2022.

[26] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. arXiv preprint arXiv:2203.02172, 2022.

[27] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5206–5215, 2022.

[28] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Testtime prompt tuning for zero-shot generalization in visionlanguage models. arXiv preprint arXiv:2209.07511, 2022.

[29] Christian Simon, Piotr Koniusz, and Mehrtash Harandi. Meta-learning for multi-label few-shot classification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 3951–3960, 2022.

[30] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. arXiv preprint arXiv:2206.09541, 2022.