**BÁO CÁO**

**DỰ ÁN CUỘC THI KHKT HỌC SINH TRUNG HỌC CẤP CỤM**

**NĂM HỌC 2024 - 2025**

**---------------\*---------------**

**DỰ ÁN**

**Phát triển ứng dụng ghi chú thông minh sử dụng công nghệ Deep Learning**

**LĨNH VỰC: Phần mềm hệ thống**

***Hà Nội ,* *tháng 11 năm 2024***

**BÁO CÁO KẾT QUẢ NGHIÊN CỨU**

**MỤC LỤC**

[**A. MỞ ĐẦU** 2](#_gjdgxs)

**B. NỘI DUNG**

***Chương 1: Cơ sở lý thuyết***

*1.1 Deep learning và neural networks* [6](#_1t3h5sf)

*1.2 Computer Vision (Thị giác máy tính)* [6](#_4d34og8)

*1.3 MLLM (Multimodal Large Language Model* [6](#_4d34og8)

***Chương 2: Phương pháp*** [7](#_2s8eyo1)

*2.1 Thiết kế module tính toán* [7](#_17dp8vu)

*2.2 Thiết kế module xử lý ngôn ngữ đa phương tiện*  [10](#_3rdcrjn)

*2.3 Xây dựng ứng dụng hệ thống* [10](#_26in1rg)

***Chương 3: Thực nghiệm***

*3.1. Mẫu dữ liệu huấn luyện mô hình phát hiện vật thể* [*14*](#_44sinio)

*3.2. Kết quả* [*14*](#_44sinio)

**C. TỔNG KẾT DỰ ÁN**

*4.1 Kết luận* [7](#_17dp8vu)

*4.2 Hạn chế tồn tại* [7](#_17dp8vu)

*4.2 Hướng phát triển trong tương lai* [7](#_17dp8vu)

**D. TÀI LIỆU THAM KHẢO**

**A. Mở đầu**

Trong kỷ nguyên công nghệ số hiện nay, nhu cầu về các ứng dụng ghi chú thông minh không ngừng gia tăng, khi con người ngày càng đòi hỏi sự tiện lợi, tốc độ và hiệu quả trong học tập lẫn công việc, đặc biệt là những công cụ hỗ trợ tính toán và xử lý thông tin tức thì. Tuy nhiên, các công cụ ghi chú truyền thống vẫn còn nhiều hạn chế: chỉ tập trung vào lưu trữ thông tin mà thiếu khả năng xử lý nhanh, nhận diện chữ viết tay thời gian thực, hay hỗ trợ tương tác liền mạch với người dùng. Hệ quả là người dùng thường phải chuyển đổi giữa nhiều công cụ riêng lẻ, làm gián đoạn dòng suy nghĩ và giảm hiệu quả làm việc.

Dự án này được ra đời và tập trung phát triển một ứng dụng tích hợp công nghệ deep learning, cho phép nhận diện chữ viết tay và thực hiện tính toán ngay trên canvas trong thời gian thực. Người dùng có thể thực hiện các phép toán một cách nhanh chóng và chính xác mà không cần phải chuyển đổi giữa các thao tác, giúp tiết kiệm thời gian và tăng hiệu quả.

Ngoài tính năng tính toán từ chữ viết tay, ứng dụng còn tích hợp tính năng trò chuyện thời gian thực, cho phép người dùng tương tác với AI không chỉ để tra cứu thông tin mà còn để nhận hướng dẫn như một gia sư, có thể giải đáp thắc mắc, giải thích các khái niệm, hoặc hỗ trợ người dùng trong quá trình học tập và giải quyết vấn đề.

Nhờ sự kết hợp giữa tính năng xử lý tức thì và khả năng giao tiếp tự nhiên, ứng dụng hứa hẹn mang lại một công cụ đột phá, giúp người dùng tối ưu hóa quá trình ghi chú, duy trì dòng suy nghĩ mạch lạc và nâng cao hiệu suất công việc. Đây không chỉ là một bước tiến công nghệ mà còn là lời giải cho những nhu cầu hiện đại về tính linh hoạt và hiệu quả.

Tổng kết với những tính năng tiên tiến này, dự án hứa hẹn mang đến một công cụ hỗ trợ đắc lực trong học tập và công việc, đáp ứng nhu cầu hiện đại về tính tiện lợi và linh hoạt phục vụ cho người dùng trong cuộc sống hằng ngày.

**B. NỘI DUNG**

***Chương I: Cơ sở lý thuyết***

**1.1 Deep learning và neural networks**

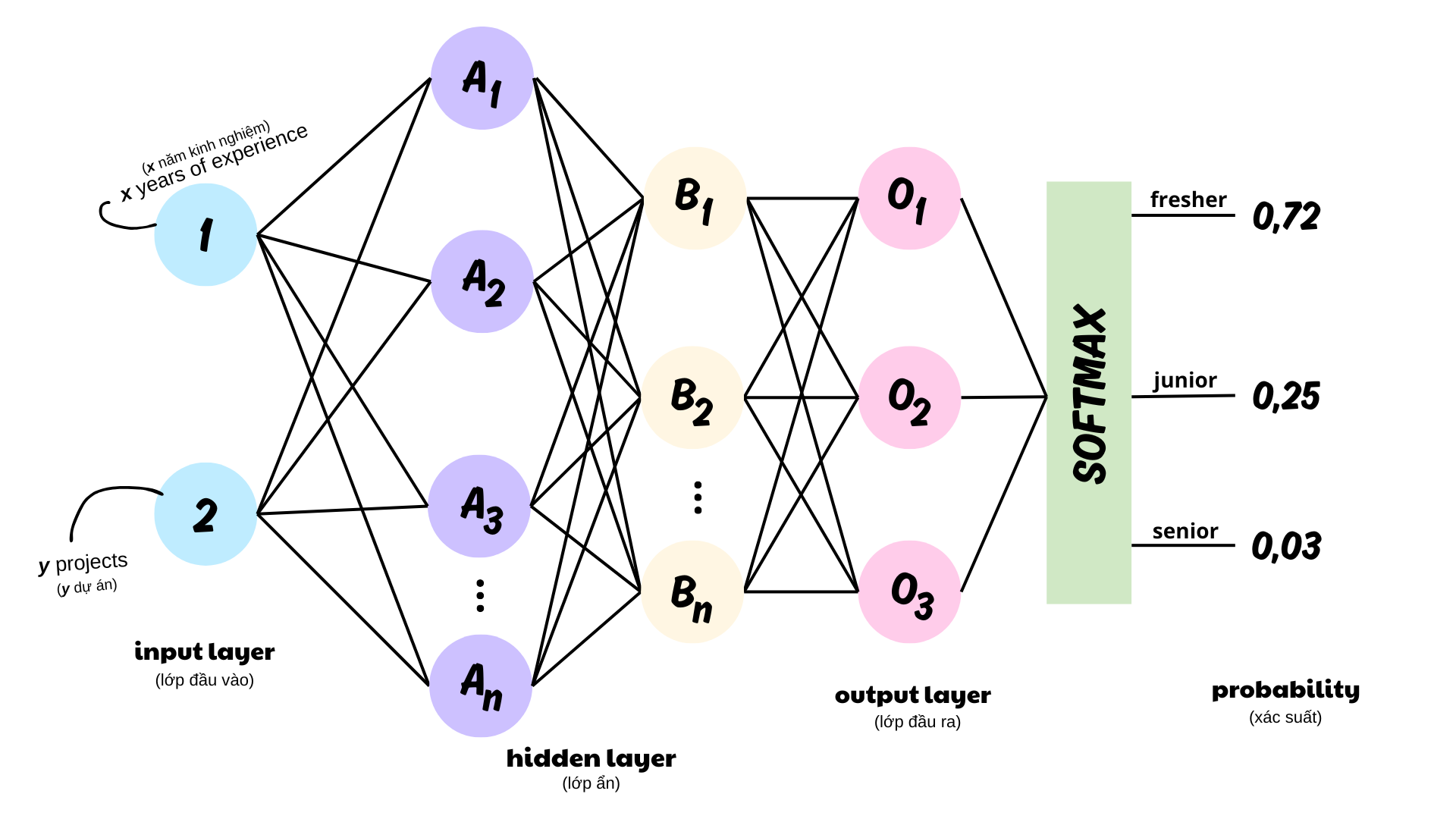
**Deep Learning (học sâu)** là một nhánh của AI, hoạt động dựa trên mạng nơ-ron nhân tạo (neural networks) để mô phỏng cách thức hoạt động của não bộ con người trong việc phân tích và xử lý dữ liệu. Với cấu trúc linh hoạt, Deep Learning có khả năng giải quyết các vấn đề phức tạp và xử lý lượng dữ liệu lớn với độ chính xác cao. Tuy nhiên, nó đòi hỏi lượng dữ liệu khổng lồ, chi phí tính toán cao, và thiếu nền tảng lý thuyết vững chắc trong việc chọn lựa công cụ tối ưu.

**Neural Networks (mạng nơ-ron nhân tạo)** là một hệ thống tính toán lấy cảm hứng từ hệ thần kinh sinh học của con người. Các mạng nơ-ron bao gồm một tập hợp các đơn vị tính toán, gọi là nơ-ron, được kết nối với nhau theo cấu trúc phân lớp:

1. **Lớp đầu vào (Input Layer):** Nhận dữ liệu ban đầu và truyền đến các lớp tiếp theo.
2. **Lớp ẩn (Hidden Layers):** Gồm một hoặc nhiều lớp, nơi các phép toán phức tạp được thực hiện để trích xuất thông tin từ dữ liệu.
3. **Lớp đầu ra (Output Layer):** Trả về kết quả cuối cùng, như dự đoán hoặc phân loại.

Mỗi nơ-ron trong mạng nhận tín hiệu đầu vào từ các nơ-ron ở lớp trước, thực hiện các phép toán dựa trên trọng số liên kết và hệ số bù (bias), sau đó áp dụng một hàm kích hoạt (activation function) để tạo đầu ra. Cấu trúc này cho phép mạng nơ-ron mô hình hóa các mối quan hệ phi tuyến tính, làm cho chúng đặc biệt mạnh mẽ trong việc xử lý các dữ liệu phức tạp như hình ảnh, âm thanh và văn bản.

Neural networks là nền tảng cốt lõi của Deep Learning, giúp hệ thống học hỏi từ dữ liệu và cải thiện hiệu suất thông qua việc điều chỉnh các trọng số trong quá trình huấn luyện.



***Hình 1: Minh họa cấu trúc của một MLP - loại neural networks đơn giản nhất***

**1.2 Computer Vision (Thị giác máy tính)**

Thị giác máy tính là một lĩnh vực trong trí tuệ nhân tạo giúp máy tính "nhìn" và hiểu hình ảnh hoặc video như con người. Máy tính sẽ phân tích hình ảnh, nhận diện các đặc trưng như hình dạng, màu sắc và chuyển động, từ đó hiểu được nội dung của hình ảnh. Để làm được điều này, máy tính sử dụng các thuật toán học sâu (deep learning), đặc biệt là các mạng nơ-ron tích chập (CNN), giúp máy tính tự động học và nhận diện các đặc trưng mà không cần sự can thiệp thủ công. Thị giác máy tính không chỉ giúp nhận diện hình ảnh một cách tổng thể mà còn có các ứng dụng cụ thể hơn, như **OCR (Nhận diện ký tự quang học)** và **Object Detection (Phát hiện đối tượng)**

. Cả hai ứng dụng này đều sử dụng các kỹ thuật trong thị giác máy tính để giải quyết những bài toán đặc thù.

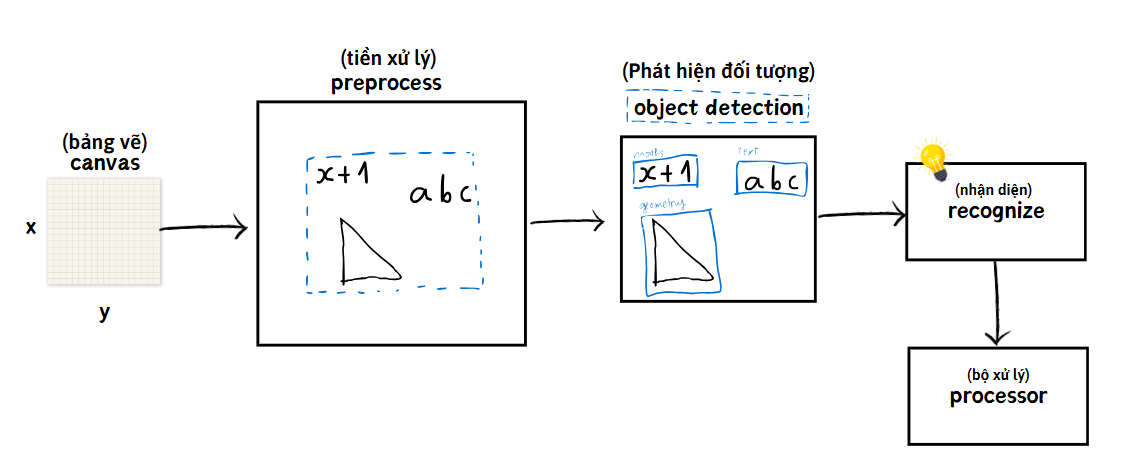
**1.3 “MLLM (Multimodal Large Language Model) Mô hình ngôn ngữ lớn đa phương thức”**

Nếu **Large Language Model (LLM)** là những mô hình học sâu được huấn luyện trên lượng lớn dữ liệu văn bản để xử lý và tạo ra ngôn ngữ tự nhiên, có thể thực hiện các nhiệm vụ như trả lời câu hỏi, dịch ngôn ngữ, tóm tắt và tạo văn bản, thì **Multimodal Large Language Model (MLLM)** là một phiên bản mở rộng, có khả năng xử lý nhiều loại dữ liệu khác nhau như văn bản, hình ảnh, âm thanh và video. Không chỉ hiểu từng loại dữ liệu riêng biệt, MLLM còn có thể kết hợp chúng để đưa ra kết quả chính xác và phù hợp với ngữ cảnh.

Ví dụ, MLLM có thể mô tả nội dung của hình ảnh, trả lời câu hỏi từ video, hoặc kết hợp thông tin từ nhiều nguồn để tạo ra phản hồi toàn diện. Các kỹ thuật tiên tiến như Vision Transformer (ViT) dùng để xử lý hình ảnh và Transformer cho văn bản, cùng với các cơ chế như **Image Projector** , cho phép MLLM tích hợp thông tin từ nhiều modal khác nhau. Những mô hình như Flamingo, LLaVA và **Lavy** đã chứng minh tiềm năng của MLLM trong các ứng dụng như trợ lý thông minh và phân tích dữ liệu đa phương tiện.

***Chương II: Phương pháp***

***2.1.Thiết kế Module tính toán***

******

* ***Hình 2: Mô tả về cấu trúc đề xuất của module tính toán***

***2.1.1 Tiền xử lý (Preprocessing) và Phát hiện đối tượng (Object Detection)***

Quy trình tiền xử lý sử dụng phương pháp phát hiện đường viền (contour detection) để xác định và cắt bỏ các vùng không chứa thông tin quan trọng, chỉ giữ lại các khu vực có đặc trưng cần thiết. Phương pháp này giúp đơn giản hóa hình ảnh và làm nổi bật các đặc trưng quan trọng trước khi thực hiện phát hiện đối tượng.

Trong bước phát hiện đối tượng, mô hình YOLO (You Only Look Once) được áp dụng để phân loại và định vị các đối tượng trên canvas, bao gồm ba nhóm chính: hình học, công thức số, và văn bản. Mô hình được huấn luyện dựa trên nhãn cụ thể cho từng loại đối tượng, giúp tập trung xử lý vào các vùng quan trọng, cải thiện hiệu suất và độ chính xác của hệ thống.

***2.1.3 Nhận diện (Recognition) và xử lý*** ***(Processor)***

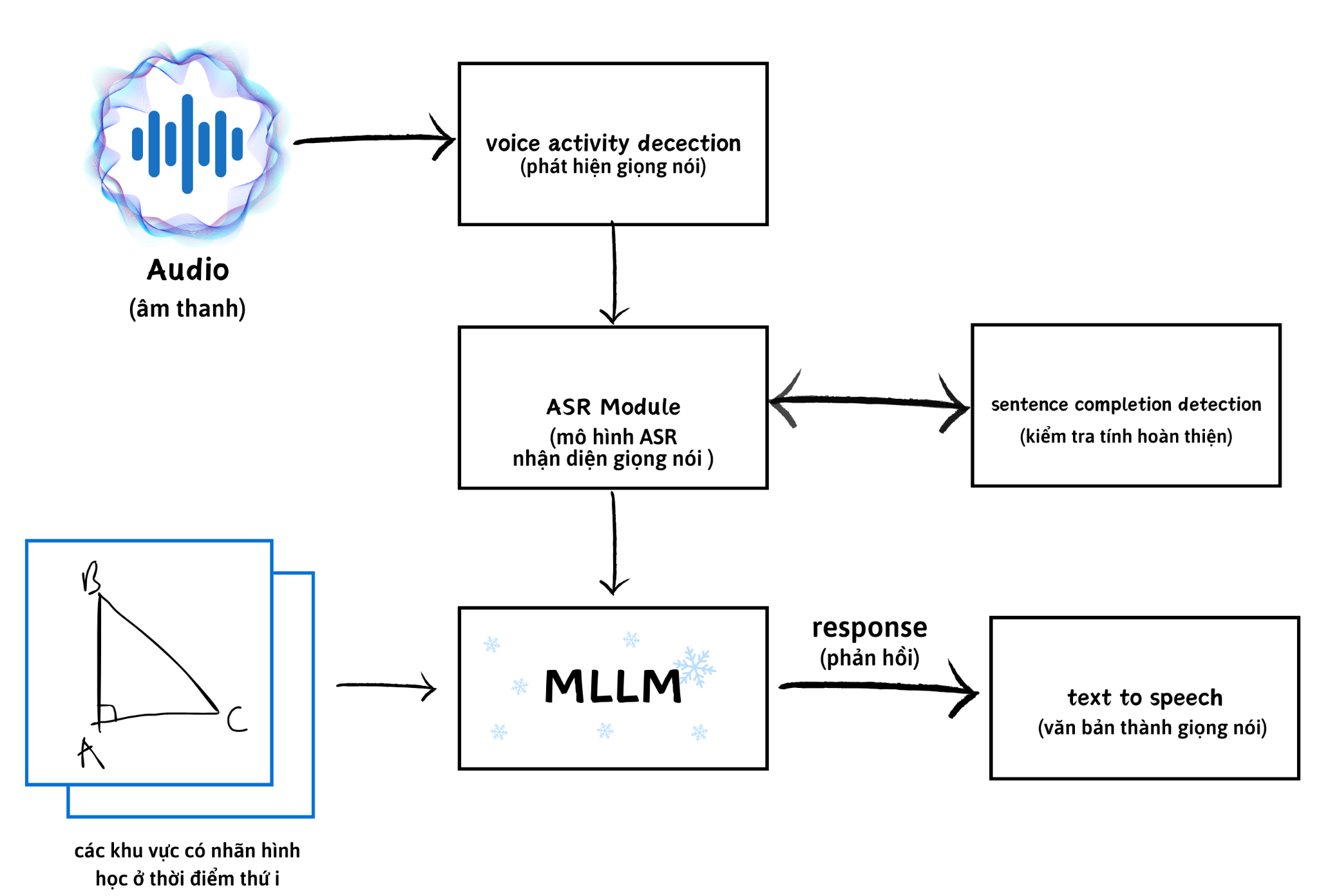
| **Đối tượng** | **Nhận diện** | **Xử lý** | **Mô hình Nhận diện được triển khai** |
| --- | --- | --- | --- |
| Hình học | Kết hợp với Văn bản và các công thức số đã nhận diện để cho vào Multimodal LLM cho phép người dùng đối thoại về nội dung trên canvas. | | Vì hạn hẹp thời gian, nhóm quyết định sử dụng GPT-4o của OpenAI để nhận diện |
| Văn bản | Sử dụng OCR để chuyển đổi thành dạng văn bản có thể chỉnh sửa và tìm kiếm được trên máy tính. | Phục vụ truy xuất, tìm kiếm trong các ghi chú với nhau |
| Công thức số | Sử dụng các thư viện mô hình toán học để có thể thực hiện các phép toán cơ bản đến phức tạp: Rút gọn biểu thức, tìm , tính toán thông thường,… | Mô hình OCR Công thức số là CoMER (Zhao and Gao) |

***Bảng 1: Cách nhận diện và xử lý các kiểu dữ liệu***

***2.2.Thiết kế Module xử lý ngôn ngữ tự nhiên đa phương tiện***

Trong bước này, module này sẽ giúp chúng ta có thể đối thoại với nội dung trong canvas cũng như là về các kiến thức liên quan. Hệ thống xử lý ngôn ngữ tự nhiên đa phương tiện bao gồm các bước chính: đầu vào âm thanh được xử lý qua module **Voice Activity Detection** để phát hiện giọng nói (đảm bảo thành phần này xác định các đoạn âm thanh có chứa giọng nói, loại bỏ khoảng lặng hoặc tiếng ồn không liên quan), sau đó được chuyển đổi từ âm thanh thành văn bản bởi **ASR Module** (**Automatic Speech Recognition -** nhận diện giọng nói tự động).

Văn bản này được kiểm tra tính hoàn thiện qua một mô hình **Sentence Completion Detection** trước khi gửi đến **MLLM** (mô hình xử lý ngôn ngữ đa phương tiện), nơi phản hồi được tạo ra. Phản hồi có thể là văn bản hoặc kết hợp với hình ảnh nhận diện từ các khu vực hình học. Cuối cùng, văn bản phản hồi được chuyển thành giọng nói qua **Text-to-Speech** để cung cấp đầu ra âm thanh, hoàn thiện luồng xử lý.

******

***Hình 2: Mô tả về module xử lý đa phương tiện***

***2.3. Xây dựng ứng dụng hệ thống***

Hệ thống sẽ bao gồm các thành phần chính sau: giao diện người dùng (UI) và server để kết nối các module với UI. Giao diện sẽ được xây dựng bằng React kết hợp với HTML5 Canvas, giúp người dùng tương tác trực tiếp với bảng trắng, vẽ và viết các công thức hay bài toán. Khi có đầu vào từ người dùng, hệ thống sẽ sử dụng module tính toán và gửi dữ liệu đến thư viện SymPy để thực hiện các phép toán từ công thức hoặc bài toán nhận diện qua viết tay. Đối với module xử lý ngôn ngữ tự nhiên, chúng tôi sẽ áp dụng mô hình GPT-4 của OpenAI để phân tích và hiểu ngữ cảnh của các đối tượng trên canvas, đồng thời tận dụng các API Text to Speech và Speech to Text của FPT AI để thực hiện các tác vụ đối thoại. Các module này sẽ giao tiếp qua API RESTful, được triển khai trên nền tảng FastAPI và ExpressJS, đảm bảo tính linh hoạt và dễ mở rộng. Hệ thống sẽ được triển khai trong môi trường Docker, giúp duy trì tính nhất quán và thuận tiện khi triển khai trên nhiều nền tảng khác nhau.

***Chương III: Thực Nghiệm***

Trong phần này, chúng tôi trình bày các kết quả thực nghiệm của mô hình đang dùng để nghiên cứu, đề xuất và so sánh đầu ra. Kết quả được đánh giá dựa trên nhiều bộ dữ liệu và các tiêu chí khác nhau, nhằm làm rõ hiệu quả cải tiến của mô hình đang dùng. Ngoài ra, chúng tôi cũng phân tích những hạn chế của mô hình và thảo luận về các hướng phát triển để khắc phục những hạn chế này trong tương lai

1. **Mẫu dữ liệu huấn luyện mô hình phát hiện vật thể**

|  |  |
| --- | --- |

***Bảng 2: Các mẫu ảnh sinh ra từ thuật toán “xào” ảnh***

| **Phép biến đổi** | **Mô tả** | **Đối tượng áp dụng** | **Tỉ lệ** |
| --- | --- | --- | --- |
| Giãn nở, Xói mòn,  Phóng to, thu nhỏ | Làm các nét chữ, nét vẽ tăng, giảm về độ rộng. Thay đổi kích thước đối tượng | Ảnh chữ viết, công thức số, hình học | 50% cho mỗi phép trên mỗi ảnh phần tử. |
| Xoay, lật, cắt bỏ | Thay đổi góc độ, phương hướng của đối tượng, chiều ngẫu nhiên. | Hình học |
| Tạo nét cong, đường thẳng, chấm tròn | Thêm các yếu tố hình học cơ bản như đường thẳng, đường cong Bezier, chấm tròn ngẫu nhiên vào hình. | Toàn cục | 20% cho mỗi phép biến đổi |

**Bảng 3. Các phép biến đổi ảnh**

Do tính đặc thù của bài toán, chúng tôi sẽ sử dụng các hình ảnh từ datasets ảnh từ internet cũng như những ảnh nội bộ rồi tạo một canvas và một thuật toán để có thể đặt danh sách các ảnh với mỗi ảnh có tại một vị trí ngẫu nhiên thỏa mãn các điều kiện sau:

Tức là các ảnh không được nằm ngoài canvas (1) cũng như không va chạm với nhau (2). Bằng cách này kết hợp với các phép biến đổi ảnh ở **Bảng 3**, chúng tôi đã tạo ra một bộ dữ liệu đa dạng để có thể huấn luyện mô hình phát hiện vật thể một cách hiệu quả. Tránh các trường hợp mô hình học các đặc điểm không mang tính tổng quát.

**2. Kết quả**

Sản phẩm của chúng tôi được triển khai tại [đây](https://www.thegioididong.com/may-tinh-bang-samsung).

Còn dưới đây là kết quả của quá trình huấn luyện mô hình phát hiện:

| **Epoch** | **Params** | **AP50** | **AP50-95** |
| --- | --- | --- | --- |
| 15 | 20.1M | 96.7 | 87.3 |

***Bảng 4:Kết quả của model được trên bộ dữ liệu kiểm thử***

Dù bảng hiệu suất thể hiện mô hình có thể phát hiện các vùng với độ chính xác gần như tuyệt đối - dù là trên bộ dữ liệu để kiểm thử. Cần lưu ý rằng tất cả dữ liệu này được tạo ra từ chung một thuật toán. Vậy nên không tránh khỏi mô hình đã làm quen với những đặc điểm sẵn có. Do đó, sẽ là hợp lý để tiến hành thử nghiệm trên những ảnh độc lập.

| **Ảnh đầu vào** | **Các vùng mô hình đã phát hiện và phân loại** |
| --- | --- |
|  |  |
|  |  |

***Bảng 5: Kiểm thử mô hình với những mẫu ảnh không được tạo thành từ thuật toán***

**C. TỔNG KẾT DỰ ÁN**

***3.1 Kết luận***

Qua bài báo cáo vừa rồi, nhóm chúng tôi đã giới thiệu và phát triển thành công một ứng dụng ghi chú thông minh tích hợp công nghệ deep learning với những đóng góp chính sau:

1. Xây dựng được hệ thống nhận diện và phân loại đối tượng trên canvas với độ chính xác cao (AP50 đạt 96.7% và AP50-95 đạt 87.3% trên tập kiểm thử), cho phép phân biệt được các dạng nội dung khác nhau như văn bản, công thức số và hình học.
2. Thiết kế được module tính toán thông minh có khả năng xử lý các phép toán từ đơn giản đến phức tạp thông qua nhận diện chữ viết tay, giúp người dùng thực hiện các phép tính một cách nhanh chóng và hiệu quả.
3. Phát triển thành công module xử lý ngôn ngữ đa phương tiện (MLLM) cho phép tương tác tự nhiên với nội dung trên canvas thông qua giọng nói và văn bản, tạo ra trải nghiệm học tập và làm việc liền mạch.
4. Xây dựng được phương pháp tạo dữ liệu huấn luyện hiệu quả thông qua thuật toán sinh ảnh tự động với các phép biến đổi đa dạng, giúp tăng cường chất lượng của mô hình nhận diện.

Dự án đã chứng minh được tính khả thi và hiệu quả của việc ứng dụng các công nghệ trí tuệ nhân tạo tiên tiến vào việc phát triển công cụ ghi chú thông minh. Sản phẩm không chỉ giải quyết được các hạn chế của các công cụ ghi chú truyền thống mà còn mở ra những khả năng mới trong việc tương tác với nội dung học tập và làm việc.

***3.2 Hạn chế***

Dù vậy dự án vẫn còn những hạn chế

* Chi phí triển khai khá cao, yêu cầu các phần cứng tầm trung.
* Tốc độ còn chưa được tối ưu
* Mô hình phát hiện chưa được huấn luyện trên các cấu trúc phức tạp

***3.3 Hướng phát triển***

Trong tương lai nhóm chúng tôi sẽ tập chung vào các hướng phát triển sau

* Khắc phục những hạn chế còn tồn tại
* Nhận diện chữ viết tay bằng mạng hồi tiếp dài ngắn (LSTM) và Phân loại thời gian kết nối (CTC) bằng cách tiếp nhận danh sách các điểm cấu tạo nên nét chữ đó. Giúp giảm thiệu thời gian cũng như chi chi phí để nhận diện chữ viết tay một cách đáng kể.
* Huấn luyện một mô hình riêng biệt để mô tả hình vẽ giúp gia tăng độ chính xác cho hệ thống.
* Hoàn thiện các tính năng cơ bản của một ứng dụng ghi chú.
* Cho phép người dùng vẽ trực tiếp lên file PDF, DOCX để nâng cao tính tiện lợi của ứng dụng.
* Phát triển tính năng tìm kiếm, truy xuất dữ liệu.

**D. TÀI LIỆU THAM KHẢO**

Zhao, Wenqi, and Liangcai Gao. “[2207.04410] CoMER: Modeling Coverage for Transformer-based Handwritten Mathematical Expression Recognition.” *arXiv*, 10 July 2022, https://arxiv.org/abs/2207.04410. Accessed 19 November 2024.