TRƯỜNG ĐẠI HỌC KIẾN TRÚC HÀ NỘI

**KHOA CÔNG NGHỆ THÔNG TIN**

**----- o0o -----**

A blue and white logo

Description automatically generated with low confidence

**ĐỒ ÁN TỐT NGHIỆP**

**Tên đề tài:**

**NGHIÊN CỨU XỬ LÝ NGÔN NGỮ TỰ NHIÊN VÀ XÂY DỰNG ỨNG DỤNG TRẢ LỜI VĂN BẢN HÀNH CHÍNH BẰNG PYTHON**

|  |  |
| --- | --- |
| **GIÁO VIÊN HƯỚNG DẪN:** | **TH.SĨ NGUYỄN THỊ HUỆ** |
| **SINH VIÊN THỰC HIỆN:** | **HOÀNG TUẤN VŨ** |
| **MÃ SINH VIÊN :**  **LỚP:** | **2055010293**  **20CN5** |

**Hà Nội, tháng 10 năm 2024**

**Lời nói đầu**

Ngày nay với sự phát triển của khoa học cộng nghệ và đặc biệt là sự bùng nổ trong lĩnh vực Trí tuệ nhân tạo, đã mang lại cho con người nhiều tiện ích trong cuộc sống. Dựa trên tiêu chí đó để có thể giúp người dùng trong lĩnh vực văn bản hành chính trong công việc và đời sống, đồ án này nhằm nghiên cứu về lĩnh vực Xử lý Ngôn ngữ Tự nhiên với mục tiêu xây dựng một ứng dụng hỗ trợ trả lời các văn bản hành chính.

Xử lý Ngôn ngữ Tự nhiên là một lĩnh vực không còn mới tuy nhiên vẫn luôn là một trong những lĩnh vực quan trọng nhất của Trí tuệ nhận tạo hiện nay. Những thành tựu trong lĩnh vực này không chỉ đơn thuần thay đổi cách con người tương tác với công nghệ mà còn góp phần thúc đẩy quá trình chuyển đổi số, mang lại những giá trị to lớn cho xã hội.

Thông qua việc nghiên cứu chi tiết về đề tài ứng dụng xử lý ngôn ngữ tự nhiên về văn bản hành chính, với mục đích để có thể hiểu hơn về kiến thức trí tuệ nhân tạo nói chung cũng như mong muốn có thể giúp đỡ được người dùng.

Em muốn bày tỏ lòng biết ơn chân thành đến cô Nguyễn Thị Huệ đã hỗ trợ và hướng dẫn em suốt quá trình thực hiện đề tài này. Sự đóng góp của cô đã giúp em hiểu rõ hơn về các vấn đề gặp phải và hoàn thiện đề tài của mình.

Mục lục

[Chương 1: Cơ sở lý thuyết 4](#_Toc183433764)

[1.1. Tổng quan về trí tuệ nhân tạo 4](#_Toc183433765)

[1.1.1. Khái niệm 4](#_Toc183433766)

[1.1.2. Các hướng nghiên cứu 5](#_Toc183433767)

[1.1.3. Phân loại trí tuệ nhân tạo 6](#_Toc183433768)

[1.1.4. Ứng dụng 8](#_Toc183433769)

[1.1.5. Mặt tích cực và hạn chế của trí tuệ nhân tạo 10](#_Toc183433770)

[1.2. Tổng quan về đề tài 11](#_Toc183433771)

[1.2.1. Yêu cầu 11](#_Toc183433772)

[1.2.2. Mô tả bài toán 12](#_Toc183433773)

[1.2.3. Mô hình 12](#_Toc183433774)

[1.2.4 Datasets 15](#_Toc183433775)

[1.3. Công nghệ sử dụng 16](#_Toc183433776)

[1.3.1 Python 16](#_Toc183433777)

[1.3.2 Transformer 17](#_Toc183433778)

[1.3.3 PyTorch 17](#_Toc183433779)

[1.3.4 Flask 17](#_Toc183433780)

[Chương 2: Quy trình 19](#_Toc183433781)

[2.1. Chuẩn bị tập dữ liệu 19](#_Toc183433782)

[2.2. Tiền xử lý dữ liệu 19](#_Toc183433783)

[2.3. Lựa chọn mô hình 19](#_Toc183433784)

[2.4. Huấn luyện mô hình 19](#_Toc183433785)

[2.5. Đánh giá kết quả 19](#_Toc183433786)

[Chương 3: Triển khai và thực nghiệm 19](#_Toc183433787)

[3.1 Công cụ sử dụng 19](#_Toc183433788)

[3.2 Phân tích thiết kế hệ thống 19](#_Toc183433789)

[3.3 Quy trình thực nghiệm 19](#_Toc183433790)

[3.4 Đánh giá mô hình 19](#_Toc183433791)

[3.5 Kết luận 19](#_Toc183433792)

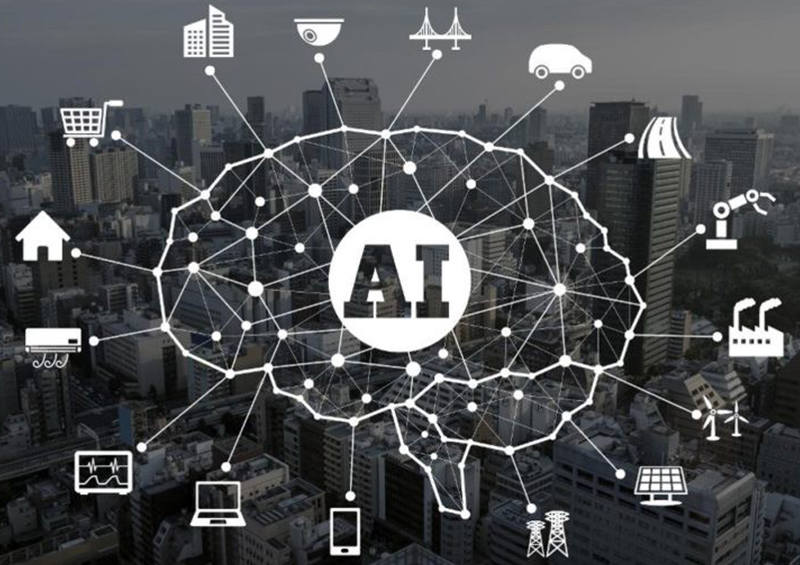
Mục lục hình ảnh và ký hiệu viết tắt

# Chương 1: Cơ sở lý thuyết

## 1.1. Tổng quan về trí tuệ nhân tạo

### 1.1.1. Khái niệm

Trí tuệ nhân tạo (AI) là một lĩnh vực trong khoa học máy tính và công nghệ thông tin tập trung vào việc phát triển máy tính và hệ thống có khả năng thực hiện các nhiệm vụ thông minh mà trước đây chỉ có con người có thể thực hiện. Mục tiêu của AI là tạo ra các chương trình máy tính hoặc hệ thống thông minh, cho phép chúng học hỏi, tự điều chỉnh và thực hiện các tác vụ yêu cầu sự hiểu biết, lý thuyết cũng như khả năng giải quyết vấn đề.



Hình 1.1 Trí tuệ nhân tạo

### 1.1.2. Các hướng nghiên cứu

Trí tuệ nhân tạo (AI) là một lĩnh vực đa dạng và rất rộng với nhiều hướng nghiên cứu khác nhau mà các nhà nghiên cứu trên khắp thế giới đang tiến hành. Dưới đây là một số hướng nghiên cứu quan trọng trong lĩnh vực này:

* Học máy và Học sâu (Machine Learning and Deep Learning): Nghiên cứu về các thuật toán học máy và học sâu để cải thiện khả năng học và tự điều chỉnh của các hệ thống AI. Điều này bao gồm việc tối ưu hóa các mô hình học máy hiện có và phát triển các mô hình mới.
* Thị giác máy tính (Computer Vision): Nghiên cứu về khả năng của máy tính để nhận diện, hiểu và xử lý hình ảnh và video. Các ứng dụng bao gồm nhận diện khuôn mặt, phân tích hình ảnh y tế và xe tự hành.
* Xử lý ngôn ngữ tự nhiên (NLP): Nghiên cứu về khả năng của máy tính để hiểu và tạo ra ngôn ngữ tự nhiên. Điều này liên quan đến chatbots, dịch máy, phân tích cảm xúc và nhiều ứng dụng khác.
* Trí Tuệ Nhân Tạo Đàm Thoại (Conversational AI): Là 1 nhánh của xử lý ngôn ngữ tự nhiên có khả năng phân tích và hiểu giọng nói của con người. Có thể giao tiếp phát âm bằng giọng nói giống như con người.
* Robotics và Trí tuệ nhân tạo trong Robot (Robotics and AI in Robotics): tập trung vào phát triển robot có khả năng tương tác với môi trường xung quanh một cách tự động. Các robot có thể thực hiện nhiệm vụ từ đơn giản đến phức tạp, như trong công nghiệp, y tế và dịch vụ khách hàng.
* Hệ thống thông minh: Hệ thống thông minh là các hệ thống có khả năng xử lý dữ liệu phức tạp, học hỏi từ kinh nghiệm và đưa ra quyết định hoặc hành động thông minh. Chúng thường kết hợp các kỹ thuật từ nhiều lĩnh vực AI.

### 1.1.3. Phân loại trí tuệ nhân tạo

Bốn loại của trí tuê nhân tạo đó là: Máy phản ứng, Bộ nhớ hạn chế, Lý thuyết tâm trí, Tự nhận thức.

* Máy phản ứng (Reactive Machines):
  + Máy phản ứng là cấp độ đơn giản nhất của AI. AI sẽ có khả năng phân tích những động thái khả nghi nhất của mình và đối thủ. Sau đó, sẽ đưa ra giải pháp tốt nhất.
  + Ví dụ: Deep Blue của IBM, một cỗ máy được thiết kế để chơi cờ vua với con người. Deep Blue đánh giá các quân cờ trên bàn cờ và phản ứng với chúng, dựa trên các chiến lược cờ vua được mã hóa trước. Deep Blue không học hỏi hoặc cải thiện khi chơi – nó chỉ đơn giản là ‘phản ứng’. Và nó đánh bại kiện tướng cờ vua Garry Kasparov vào năm 1997.



Hình 1.2. Deep Blue thi đấu cùng với Karparov

* Bộ nhớ hạn chế (Limited Memory):
  + Máy có bộ nhớ hạn chế, có thể giữ lại một số thông tin học được từ việc quan sát các sự kiện hoặc dữ liệu trước đó. AI có thể xây dựng kiến thức bằng cách sử dụng bộ nhớ đó kết hợp với dữ liệu được lập trình sẵn.
  + Ví dụ: Chatbot Thông Minh ghi nhớ câu hỏi trước của người dùng trong cuộc trò chuyện từ đó đưa ra phản hồi tốt hơn.
* Lý thuyết tâm trí (Theory of Mind):
  + Con người có những suy nghĩ và cảm xúc, ký ức hoặc các mô hình não khác điều khiển và ảnh hưởng đến hành vi của họ.
  + Dựa trên tâm lý này, các nhà nghiên cứu lý thuyết về tâm trí hy vọng phát triển các máy tính có khả năng bắt chước các mô hình tinh thần của con người. Máy móc có thể hiểu rằng con người và động vật có những suy nghĩ và cảm xúc có thể ảnh hưởng đến hành vi của chính chúng.
  + Lý thuyết về máy móc tâm trí sẽ được yêu cầu sử dụng thông tin thu được từ con người và học hỏi từ nó, sau đó sẽ thông báo bằng cách máy móc giao tiếp hoặc phản ứng với một tình huống khác
  + Ví dụ: AI hỗ trợ bác sĩ trong việc đánh giá trạng thái cảm xúc của bệnh nhân, đặc biệt trong các lĩnh vực như tâm lý học hoặc trị liệu.
* Tự nhận thức (Self-awareness):
  + Công nghệ AI này có khả năng tự nhận thức về bản thân, có ý thức và hành xử như con người. Thậm chí, chúng còn có thể bộc lộ cảm xúc cũng như hiểu được những cảm xúc của con người. Đây được xem là bước phát triển cao nhất của công nghệ AI và đến thời điểm hiện tại, công nghệ này vẫn chưa khả thi.

### 1.1.4. Ứng dụng

* Xử lý ngôn ngữ tự nhiên (NLP):
  + Chatbots và Virtual Assistants: AI được sử dụng để xây dựng các trợ lý ảo và chatbot để tương tác với người dùng và cung cấp thông tin hoặc giải quyết vấn đề.
  + Phân tích tình cảm: AI có thể phân tích tình cảm trong văn bản, giúp doanh nghiệp hiểu đánh giá của khách hàng về sản phẩm hoặc dịch vụ.
* Thị giác máy tính (Computer Vision):
  + Nhận diện khuôn mặt: AI có thể nhận diện và xác định khuôn mặt trong hình ảnh hoặc video, thường được sử dụng trong hệ thống đảm bảo an ninh và xác thực.
  + Xử lý hình ảnh y tế: Giúp bác sĩ chẩn đoán bệnh dựa trên hình ảnh y tế như X-quang, MRI, CT scans.
* Tự động hóa và Robotica:
  + Ô tô tự lái: AI được sử dụng trong xe ô tô tự lái để phát hiện và tránh các vật thể, biển báo giao thông, và điều khiển xe.
  + Robot công nghiệp: Robot được trang bị trí tuệ nhân tạo để thực hiện các nhiệm vụ trong môi trường sản xuất, từ lắp ráp đến kiểm tra chất lượng.
* Dự đoán Tài chính:
  + Dự đoán thị trường tài chính: AI được sử dụng để dự đoán xu hướng thị trường chứng khoán và thực hiện giao dịch tự động.
  + Quản lý rủi ro: Các công ty bảo hiểm sử dụng AI để định giá và quản lý rủi ro.
* Y tế và Chăm sóc sức khỏe:
  + Chẩn đoán bệnh: AI có thể phân tích dữ liệu y tế và hình ảnh để hỗ trợ bác sĩ trong việc chẩn đoán và điều trị bệnh.
  + Quản lý dữ liệu bệnh nhân: Hệ thống AI có thể quản lý và xử lý thông tin liên quan đến bệnh nhân, giúp cải thiện chất lượng dịch vụ y tế.
* Giáo dục:
  + Học tập trực tuyến và cá nhân hóa giáo dục: AI có thể tạo ra nội dung giảng dạy cá nhân hóa dựa trên nhu cầu và khả năng của từng học sinh.
  + Đánh giá học sinh: Hệ thống AI có thể phân tích kết quả học tập và đề xuất cải thiện cho học sinh.
* Điều khiển và Tự động hóa nhà thông minh:
  + Quản lý năng lượng: Hệ thống AI có thể điều khiển thiết bị để tiết kiệm năng lượng trong nhà thông minh.
  + Bảo mật: AI có thể giúp giám sát và bảo vệ nhà thông minh khỏi các mối đe dọa an ninh.
* Quảng cáo và Tiếp thị:
  + Tối ưu hóa quảng cáo trực tuyến: AI giúp tối ưu hóa chiến dịch quảng cáo trực tuyến dựa trên dữ liệu về tương tác của người dùng.
* Khám phá dữ liệu và Nghiên cứu khoa học:
  + Xử lý dữ liệu lớn (Big Data): AI giúp phân tích và rút ra thông tin quan trọng từ các nguồn dữ liệu lớn và phức tạp.

### 1.1.5. Mặt tích cực và hạn chế của trí tuệ nhân tạo

Trí tuệ nhân tạo (AI) mang lại nhiều lợi ích và tiềm năng cách mạng hóa nhiều lĩnh vực khác nhau. Tuy nhiên, cũng có những tích cực và hạn chế cần xem xét:

Tích cực:

* Tăng năng suất và hiệu suất công việc: AI có thể thực hiện nhiều nhiệm vụ một cách nhanh chóng và chính xác, giúp tăng năng suất và hiệu suất làm việc trong các ngành công nghiệp.
* Dự đoán và ứng dụng trong quản lý tài chính: AI có khả năng phân tích dữ liệu tài chính phức tạp và dự đoán xu hướng thị trường, giúp đưa ra quyết định tài chính hiệu quả.
* Quản lý y tế và chẩn đoán bệnh: Trong lĩnh vực y tế, AI có thể hỗ trợ các bác sĩ trong việc chẩn đoán bệnh và lập kế hoạch điều trị dựa trên phân tích hình ảnh và dữ liệu bệnh lý.
* Ô tô tự hành: AI có tiềm năng thúc đẩy phát triển xe tự hành, giảm tai nạn giao thông và giúp người già hoặc khuyết tật di chuyển dễ dàng hơn.
* Dự đoán thời tiết và khí hậu: AI có khả năng xử lý dữ liệu khí hậu lớn và dự đoán thời tiết một cách chính xác, giúp cảnh báo thiên tai và khắc phục hậu quả.
* Hỗ trợ trong giáo dục: AI có thể cung cấp các công cụ học tập thông minh, cá nhân hóa giáo dục và tạo điều kiện học tập tốt hơn cho học sinh.

Hạn chế:

* Sự lo ngại về đạo đức và quyền riêng tư: Sử dụng AI có thể dẫn đến các vấn đề về quyền riêng tư và đạo đức, bao gồm việc thu thập và sử dụng dữ liệu cá nhân một cách không đúng mục đích.
* Thất nghiệp và thay thế công việc: Các hệ thống tự động hóa dự kiến sẽ thay thế một số công việc, gây ra lo ngại về thất nghiệp và sự bất ổn kinh tế.
* Giới hạn trong việc hiểu và giải quyết bài toán phức tạp: AI hiện tại vẫn gặp khó khăn trong việc hiểu và giải quyết bài toán phức tạp mà con người có thể làm.
* Nguy cơ trục trặc và lỗi hệ thống: AI có thể gặp trục trặc và lỗi hệ thống, đặc biệt khi dựa vào dữ liệu không chính xác hoặc không đủ lớn.
* Khả năng phân biệt đạo đức và quyết định etic: AI hiện tại không có khả năng phân biệt đạo đức và quyết định theo tiêu chuẩn etic, gây ra các vấn đề đạo đức liên quan đến quyết định của chúng.
* Phụ thuộc vào dữ liệu lớn: AI cần dữ liệu lớn để hoạt động hiệu quả, và điều này có thể tạo ra vấn đề về riêng tư và bảo mật dữ liệu.

## 1.2. Tổng quan về đề tài

### 1.2.1. Yêu cầu

Ứng dụng trả lời văn bản trả lời được những câu hỏi:

* Tên của văn bản hành chính
* Nội dung của văn bản hành chính
* Ngày ban hành, cơ quan ban hành của văn bản hành chính

### 1.2.2. Mô tả bài toán

Lĩnh vực: Xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)

Xử lý ngôn ngữ tự nhiên là lĩnh vực trong trí tuệ nhân tạo (AI) nhằm giúp máy tính hiểu, xử lý và tương tác với ngôn ngữ tự nhiên của con người. Trong bài toán này, tập trung vào bài toán Hỏi - Đáp (Question Answering - QA).

Đầu vào (Input)

* Context: Văn bản chứa nội dung thông tin cần thiết, có thể là một đoạn văn, tài liệu, văn bản hành chính hoặc file chứa đoạn văn ví dụ như .docx .pdf .txt
* Question: Một câu hỏi được đặt ra liên quan đến nội dung trong đoạn văn bản (Context).

Đầu ra (Output)

* Answer: Kết quả trả lời câu hỏi dựa trên nội dung trong đoạn văn bản (Context). Đầu ra là một đoạn văn ngắn lấy từ Context (Extractive QA).

### 1.2.3. Mô hình

Để giải quyết bài toán Hỏi - Đáp (Question Answering), đồ án sử dụng mô hình học sâu dựa trên kiến trúc Transformer, trong đó triển khai cụ thể với phiên bản BERT được tùy chỉnh cho tiếng Việt là PhoBERT.

1.2.3.1. Kiến trúc Transformer

Transformer là một kiến trúc mô hình học sâu ra đời từ bài báo “Attention is All You Need” của Vaswani và cộng sự (2017). Transformer đã cách mạng hóa lĩnh vực Xử lý ngôn ngữ tự nhiên (NLP) nhờ khả năng xử lý dữ liệu theo cách song song và tận dụng hiệu quả các mối quan hệ ngữ nghĩa trong văn bản.

Các thành phần chính của Transformer:

* Encoder và Decoder: Transformer bao gồm hai phần chính: Encoder và Decoder. Tuy nhiên, trong bài toán Hỏi - Đáp, chỉ phần Encoder được sử dụng để hiểu ngữ cảnh và câu hỏi.
* Self-Attention Mechanism: Là thành phần cốt lõi của Transformer, cho phép mô hình tập trung vào các từ có liên quan trong ngữ cảnh, bất kể khoảng cách của chúng. Self-attention tính toán một trọng số cho từng từ trong ngữ cảnh dựa trên tầm quan trọng của nó với từ hiện tại.
* Feed-Forward Network (FFN): Sau khi tính toán Self-Attention, dữ liệu được đưa qua một mạng nơ-ron để học thêm các đặc trưng phức tạp.
* Positional Encoding: Vì Transformer không sử dụng cấu trúc tuần tự như RNN, Positional Encoding được thêm vào để cung cấp thông tin về vị trí của các từ trong câu.

Ưu điểm của Transformer:

* Khả năng xử lý song song, tăng tốc độ huấn luyện.
* Hiệu quả trong việc học mối quan hệ giữa các từ xa nhau trong văn bản dài.

1.2.3.2. BERT (Bidirectional Encoder Representations from Transformers)

BERT là một mô hình dựa trên kiến trúc Transformer, được giới thiệu bởi Google vào năm 2018. Điểm nổi bật của BERT là khả năng học ngữ cảnh hai chiều (bidirectional), cho phép hiểu ý nghĩa của một từ dựa trên cả phía trước và phía sau trong câu.

Đặc điểm của BERT:

* Masked Language Modeling (MLM): BERT được huấn luyện bằng cách dự đoán các từ bị che đi (masked) trong văn bản, giúp mô hình hiểu được ngữ nghĩa của từ dựa trên ngữ cảnh xung quanh.
* Next Sentence Prediction (NSP): Một nhiệm vụ khác trong quá trình huấn luyện BERT là dự đoán câu tiếp theo có phù hợp với câu hiện tại hay không, giúp mô hình hiểu được quan hệ giữa các câu.

### 1.2.4 Datasets

Tập dữ liệu sử dụng trong bài toán được xây dựng và tinh chỉnh dựa trên cấu trúc tương tự bộ dữ liệu SQuAD (Stanford Question Answering Dataset), bao gồm các trường thông tin:

Cấu trúc dữ liệu:

id: ID duy nhất cho mỗi mẫu dữ liệu.

context: Nội dung văn bản chứa thông tin cần thiết.

question: Câu hỏi liên quan đến nội dung văn bản.

answer: Câu trả lời phù hợp với câu hỏi, được lấy từ Context (hoặc tạo mới).

Nguồn dữ liệu:

* Văn bản hành chính: Các tài liệu công khai như quyết định, quy định hoặc công văn từ các cơ quan nhà nước.
* Bộ dữ liệu tiếng Việt công khai: Ví dụ như UIT-ViQuAD hoặc các tài liệu từ kho dữ liệu mở.
* Dữ liệu tự xây dựng: Tự tạo dữ liệu bằng cách tổng hợp và gán nhãn các cặp câu hỏi - câu trả lời phù hợp với ngữ cảnh.

Kích thước và chất lượng dữ liệu:

* Tập dữ liệu ban đầu bao gồm 1,000 mẫu dữ liệu được gán nhãn thủ công.
* Sử dụng phương pháp đánh giá chéo (cross-validation) để đảm bảo chất lượng và độ chính xác của dữ liệu.

Tiền xử lý dữ liệu:

* Loại bỏ các đoạn văn bản dư thừa hoặc chứa nhiều lỗi chính tả.
* Chuẩn hóa văn bản bằng cách loại bỏ ký tự đặc biệt và chuẩn hóa dấu câu.
* Mã hóa dữ liệu bằng tokenizer của Phobert để đưa vào mô hình.

## 1.3. Công nghệ sử dụng

Để giải quyết bài toán Hỏi - Đáp (Question Answering) một cách hiệu quả, đồ án sử dụng các công nghệ tiên tiến trong lĩnh vực xử lý ngôn ngữ tự nhiên (NLP) và phát triển ứng dụng. Các công nghệ được chọn bao gồm ngôn ngữ lập trình Python, thư viện học sâu, và framework xây dựng API.

### 1.3.1. Python

Python là ngôn ngữ lập trình chính được sử dụng trong đồ án nhờ vào tính đơn giản, khả năng mở rộng, và hệ sinh thái phong phú của các thư viện hỗ trợ. Python đóng vai trò quan trọng trong:

* Xử lý dữ liệu: Tiền xử lý văn bản, mã hóa đầu vào, và phân tích kết quả.
* Phát triển mô hình: Sử dụng các thư viện học sâu như PyTorch để xây dựng và huấn luyện mô hình.
* Tích hợp ứng dụng: Xây dựng giao diện lập trình ứng dụng (API) bằng Flask để triển khai mô hình trên nền tảng web.

### 1.3.2. Transformer

Transformer là kiến trúc mô hình học sâu hiện đại, đã cách mạng hóa lĩnh vực NLP. Trong đồ án, kiến trúc transformer được ứng dụng thông qua mô hình Phobert (vinai/phobert-base-v2), được tinh chỉnh để giải quyết bài toán Hỏi - Đáp.

Các đặc điểm nổi bật:

* Self-Attention Mechanism: Giúp mô hình nắm bắt mối quan hệ giữa các từ trong ngữ cảnh một cách hiệu quả, ngay cả với các đoạn văn bản dài.
* Fine-Tuning: Tinh chỉnh mô hình transformer để thích ứng với bài toán cụ thể, giúp cải thiện độ chính xác trong việc trả lời câu hỏi.
* Tokenizer: Sử dụng cơ chế Byte-Pair Encoding (BPE) được tối ưu hóa cho ngôn ngữ tiếng Việt, hỗ trợ việc xử lý văn bản hiệu quả.

### 1.3.3. PyTorch

PyTorch là thư viện mã nguồn mở mạnh mẽ cho học sâu, được sử dụng để xây dựng và huấn luyện mô hình trong đồ án.

Lý do chọn PyTorch:

* Giao diện thân thiện, dễ dàng kiểm tra và gỡ lỗi.
* Hỗ trợ tính toán trên GPU, tăng tốc quá trình huấn luyện mô hình.
* Tích hợp tốt với các thư viện transformer như Hugging Face Transformers.
* Ứng dụng trong đồ án:
* Định nghĩa kiến trúc mô hình dựa trên transformer.
* Xử lý và tối ưu hóa loss function trong quá trình huấn luyện.
* Đánh giá hiệu suất mô hình bằng các thước đo tiêu chuẩn.

### 1.3.4. Flask

Flask là một micro-framework Python phổ biến, được sử dụng để xây dựng giao diện lập trình ứng dụng (API) cho mô hình.

Vai trò:

* Kết nối giữa mô hình AI và người dùng qua giao diện web.
* Cung cấp các endpoint RESTful cho việc gửi câu hỏi và nhận câu trả lời.
* Đảm bảo ứng dụng hoạt động nhẹ, nhanh và dễ triển khai.

Chức năng:

* Endpoint /predict: Nhận đầu vào gồm Context và Question, sau đó trả về Answer được mô hình tạo ra.

Kiến trúc triển khai:

* Mô hình được huấn luyện và lưu dưới dạng checkpoint.
* Flask API được tích hợp để tải checkpoint và xử lý yêu cầu từ người dùng.
* Ứng dụng Flask có thể triển khai trên các nền tảng như Heroku, Vercel, hoặc máy chủ nội bộ.

# Chương 2: Quy trình

## 2.1. Chuẩn bị tập dữ liệu

Để huấn luyện mô hình cần chuẩn bị một tập dữ liệu, tập dữ liệu là yếu tố vô cùng quan trọng trong việc huấn luyện mô hình. Tập dữ liệu sạch, đầy đủ và chất lượng sẽ giúp xây dựng mô hình AI chính xác và hiệu quả.

Để xây dựng một tập dữ liệu thống nhất và hiệu quả, tất cả các dữ liệu được lưu dưới dạng SquAD 2.0 (Standford Question Answering Dataset 2.0) là một bộ dữ liệu trong lĩnh vực xử lý ngôn ngữ tự nhiên. SquAD 2.0 là phiên bản mở rộng của SquAD 1.1, bổ sung các tính năng quan trọng nhằm giải quyết nhiều khía cạnh thực tế hơn trong việc hiểu văn bản và trả lời câu hỏi.

## 2.2. Tiền xử lý dữ liệu

## 2.3. Lựa chọn mô hình

PhoBERT: Phiên bản BERT cho tiếng Việt

PhoBERT là một mô hình transformer được phát triển dựa trên kiến trúc BERT, nhưng được tối ưu hóa cho tiếng Việt bởi VinAI. Điểm nổi bật:

* Tối ưu hóa ngôn ngữ: PhoBERT được huấn luyện trên kho dữ liệu văn bản tiếng Việt lớn, giúp mô hình hiểu sâu ngữ pháp và ngữ nghĩa đặc thù của tiếng Việt.
* Tokenization: Sử dụng kỹ thuật Byte-Pair Encoding (BPE) để chia từ thành các token nhỏ hơn, phù hợp với đặc thù ngôn ngữ đơn tiết và dấu của tiếng Việt.

Ứng dụng PhoBERT trong bài toán Hỏi - Đáp:

Fine-Tuning:

PhoBERT được tinh chỉnh trên tập dữ liệu dạng SQuAD (Stanford Question Answering Dataset) bằng cách tối ưu hóa dự đoán vị trí bắt đầu và kết thúc của câu trả lời trong ngữ cảnh.

Đầu vào của PhoBERT là tổ hợp của:

[CLS] + Question + [SEP] + Context + [SEP].

Token [CLS] chứa thông tin tổng hợp của toàn bộ đầu vào.

Output:

PhoBERT dự đoán hai điểm quan trọng trong văn bản: vị trí bắt đầu (start position) và vị trí kết thúc (end position) của câu trả lời trong đoạn văn.

Quy trình ứng dụng:

* Tiền xử lý: Token hóa Context và Question bằng tokenizer của PhoBERT. Chuyển đổi dữ liệu đầu vào thành định dạng phù hợp với mô hình.
* Huấn luyện: Sử dụng tập dữ liệu huấn luyện với thông tin về câu hỏi và câu trả lời. Tối ưu hóa hàm mất mát dựa trên vị trí bắt đầu và kết thúc của câu trả lời.
* Dự đoán: Với câu hỏi mới, mô hình trả về đoạn văn chứa câu trả lời trong Context.

## 2.4. Huấn luyện mô hình

## 2.5. Đánh giá kết quả

# Chương 3: Triển khai và thực nghiệm

## 3.1 Công cụ sử dụng

## 3.2 Phân tích thiết kế hệ thống

## 3.3 Quy trình thực nghiệm

## 3.4 Đánh giá mô hình

## 3.5 Kết luận

TÀI LIỆU THAM KHẢO

[1] Giáo trình Nhập môn Trí tuệ nhân tạo – Thầy Từ Minh Phương

[2] https://docs.python.org/3

[3] https://www.pygame.org/docs

36

[question\_answering\_bartpho\_phobert/code.ipynb at main · phkhanhtrinh23/question\_answering\_bartpho\_phobert](https://github.com/phkhanhtrinh23/question_answering_bartpho_phobert/blob/main/code.ipynb)

[abaoxomtieu/RAG-API-VietnamesePDF: Build API for Q&A on private PDF.](https://github.com/abaoxomtieu/RAG-API-VietnamesePDF)

[Báo cáo bài tập lớn - TRƯỜNG ĐẠI HỌC KIẾN TRÚC HÀ NỘI KHOA CÔNG NGHỆ THÔNG TIN ĐỒ ÁN MÔN TRÍ TUỆ - Studocu](https://www.studocu.com/vn/document/truong-dai-hoc-kien-truc-ha-noi/tri-tue-nhan-tao/bao-cao-bai-tap-lon/74287674)