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 2 năm 2025**

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 2 năm 2025**

**Lời nói đầu**

Ngày nay với sự phát triển của khoa học công 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 nhân 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** 5](#_Toc187853805)

[**1.1. Tổng quan về trí tuệ nhân tạo** 5](#_Toc187853806)

[**1.1.1. Khái niệm** 5](#_Toc187853807)

[**1.1.2. Các hướng nghiên cứu** 6](#_Toc187853808)

[**1.1.3. Phân loại trí tuệ nhân tạo** 6](#_Toc187853809)

[**1.1.4. Ứng dụng** 8](#_Toc187853810)

[**1.1.5. Mặt tích cực và hạn chế của trí tuệ nhân tạo** 10](#_Toc187853811)

[**1.2. Tổng quan về đề tài** 11](#_Toc187853812)

[**1.2.1. Yêu cầu** 11](#_Toc187853813)

[**1.2.2. Mô tả bài toán** 11](#_Toc187853814)

[**1.2.3. Mô hình** 12](#_Toc187853815)

[**1.2.4 Datasets** 17](#_Toc187853816)

[**1.3. Công nghệ sử dụng** 18](#_Toc187853817)

[**1.3.1. Python** 18](#_Toc187853818)

[**1.3.2. Transformer** 18](#_Toc187853819)

[**1.3.3. PyTorch** 19](#_Toc187853820)

[**1.3.4. Flask** 19](#_Toc187853821)

[**CHƯƠNG 2: QUY TRÌNH CHUẨN BỊ VÀ HUẤN LUYỆN MÔ HÌNH** 20](#_Toc187853822)

[**2.1. Chuẩn bị tập dữ liệu** 20](#_Toc187853823)

[**2.2. Tiền xử lý dữ liệu** 20](#_Toc187853824)

[**2.3. Phân chia dữ liệu** 20](#_Toc187853825)

[**2.4. Lựa chọn mô hình** 20](#_Toc187853826)

[**2.5. Huấn luyện mô hình** 20](#_Toc187853827)

[**2.6. Đánh giá kết quả** 21](#_Toc187853828)

[**CHƯƠNG 3: TRIỂN KHAI VÀ THỰC NGHIỆM** 24](#_Toc187853829)

[**3.1 Phân tích thiết kế hệ thống** 24](#_Toc187853830)

[**3.2 Quy trình thực nghiệm** 24](#_Toc187853831)

[**3.3 Đánh giá kết quả** 24](#_Toc187853832)

[**CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN** 24](#_Toc187853833)

[**4.1 Kết luận về đề tài** 24](#_Toc187853834)

[**4.2 Hướng phát triển của đề tài** 25](#_Toc187853835)

[**TÀI LIỆU THAM KHẢO** 26](#_Toc187853836)

[**PHỤ LỤC** 27](#_Toc187853837)

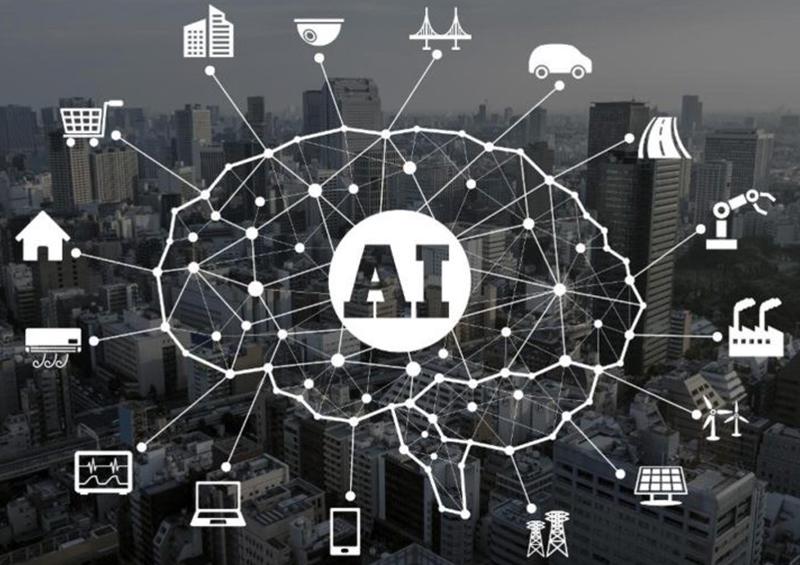
Mục lục hình ảnh và ký hiệu viết tắt

# **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 tuệ 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 Deep Learning**

## **1.3. Tổng quan xử lý ngôn ngữ tự nhiên**

## **1.4. Giới thiệu về Transformer**

## **1.5. BERT (Bidirectional Encoder Representations from Transformers)**

## **1.6. Công nghệ sử dụng**

### **1.6.1. 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.

### **1.6.2. 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.

### **1.6.3. Nuxt**

# **CHƯƠNG 2: QUY TRÌNH CHUẨN BỊ VÀ HUẤN LUYỆN MÔ HÌNH**

## **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.

Dữ liệu được thu thập là các văn bản hành chính được ban hành gồm nhiều nội dung như báo cáo, thông tư, quyết định, nghị quyết, ...

Ví dụ: Thông tư 01/2020/TT-BTP

Context: Thông tư 01/2020/TT-BTP quy định chi tiết thi hành một số điều của Luật Hộ tịch, có hiệu lực từ ngày 01/07/2020. Thông tư này áp dụng cho cơ quan hộ tịch, tổ chức và cá nhân liên quan đến đăng ký hộ tịch tại Việt Nam.

Question: Thông tư 01/2020/TT-BTP có hiệu lực từ ngày nào?

Answers: ngày 01/07/2020

## **2.2. Tiền xử lý dữ liệu**

Sau khi thu thập dữ liệu thành công cần đánh dấu vị trí của câu trả lời bắt đầu trong nội dung văn bản

Context: Thông tư 01/2020/TT-BTP quy định chi tiết thi hành một số điều của Luật Hộ tịch, có hiệu lực từ ngày 01/07/2020. Thông tư này áp dụng cho cơ quan hộ tịch, tổ chức và cá nhân liên quan đến đăng ký hộ tịch tại Việt Nam.

Question: Thông tư 01/2020/TT-BTP có hiệu lực từ ngày nào?

Answers: ngày 01/07/2020

Answer start: 96

Để mô hình có thể hiểu được ta cần phải mã hóa dữ liệu

Input IDs là một danh sách các số nguyên, trong đó mỗi số đại diện cho một token sau khi văn bản được tokenize bởi mô hình ngôn ngữ.

Vai trò của input\_ids:

* Chuyển đổi văn bản sang dạng số để mô hình có thể xử lý
* Đảm bảo mỗi từ được biểu diễn duy nhất
* Cho phép mô hình học và hiểu mối quan hệ giữa các token

Quá trình chuyển đổi diễn ra như sau:

1. Văn bản ban đầu (text) →
2. Tokenizer phân tách thành các token →
3. Mỗi token được ánh xạ tới một ID duy nhất

Vì phải có độ dài cố định nên phải thêm padding vào câu <pad>

Việc thêm padding trong xử lý ngôn ngữ tự nhiên là vô cùng quan trọng:

Đồng nhất độ dài input

* Các mô hình học máy yêu cầu input có độ dài cố định
* Các văn bản có độ dài khác nhau được điều chỉnh về một độ dài chung

Hiệu quả tính toán

* GPU và các thiết bị tính toán hoạt động hiệu quả nhất khi xử lý các tensor có kích thước đồng nhất
* Giảm thiểu overhead trong quá trình batch processing

[CLS] và [SEP] là hai token đặc biệt trong nhiều mô hình transformer như BERT, được sử dụng cho các nhiệm vụ xử lý ngôn ngữ tự nhiên.

[CLS] (Classification Token):

* Được đặt ở đầu chuỗi token
* Viết tắt của "Classification"
* Mục đích: Biểu diễn toàn bộ câu hoặc văn bản. Được sử dụng cho các nhiệm vụ phân loại văn bản. Chứa thông tin ngữ cảnh tổng thể

[SEP] (Separator Token):

* Viết tắt của "Separator"
* Được sử dụng để: Phân tách các câu hoặc các phần khác nhau của văn bản. Đánh dấu ranh giới giữa các phần trong input

Ví dụ:

Context: Thông tư 01/2020/TT-BTP quy định chi tiết thi hành một số điều của Luật Hộ tịch, có hiệu lực từ ngày 01/07/2020. Thông tư này áp dụng cho cơ quan hộ tịch, tổ chức và cá nhân liên quan đến đăng ký hộ tịch tại Việt Nam.

Question: Thông tư 01/2020/TT-BTP có hiệu lực từ ngày nào?

Answer: ngày 01/07/2020

Input IDs: [0, 4848, 1713, 7818, 3914, 5419, 12607, 253, 10, 3256, 2792, 39, 43, 40865, 1672, 114, 2, 2, 4848, 1713, 7818, 3914, 5419, 12607, 253, 4026, 2608, 811, 3365, 201, 2537, 16, 100, 184, 7, 761, 13977, 34562, 1347, 4, 10, 3256, 2792, 39, 43, 7818, 16425, 3914, 3914, 5, 4848, 1713, 23, 2925, 8410, 13, 2209, 2665, 506, 34562, 1347, 4, 1391, 1766, 6, 569, 1362, 2657, 2665, 30, 1813, 765, 506, 26865, 35, 350, 43557, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

Tokens: ['<s>', 'Thông', 'tư', '01/@@', '20@@', '20/@@', 'TT-B@@', 'TP', 'có', 'hiệu', 'lực', 'từ', 'ngày', 'nà@@', 'o@@', '?', '</s>', '</s>', 'Thông', 'tư', '01/@@', '20@@', '20/@@', 'TT-B@@', 'TP', 'quy', 'định', 'chi', 'tiết', 'thi', 'hành', 'một', 'số', 'điều', 'của', 'Luật', 'Hộ', 'tị@@', 'ch@@', ',', 'có', 'hiệu', 'lực', 'từ', 'ngày', '01/@@', '07/@@', '20@@', '20@@', '.', 'Thông', 'tư', 'này', 'áp', 'dụng', 'cho', 'cơ', 'quan', 'hộ', 'tị@@', 'ch@@', ',', 'tổ', 'chức', 'và', 'cá', 'nhân', 'liên', 'quan', 'đến', 'đăng', 'ký', 'hộ', 'tịch', 'tại', 'Việt', 'Nam.', '</s>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>', '<pad>']

## **2.3. Phân chia dữ liệu**

Sau khi có được tệp dữ liệu tiến hành chia tất cả dữ liệu đã thu thập thành các tập hợp con, quá trình này được gọi là phân tách dữ liệu. Dữ liệu được chia nhỏ thành tập dữ liệu Train - đào tạo, Validation - xác thực và Test - thử nghiệm.

Về cơ bản, có nhiều cách phân chia các tập dữ liệu khác nhau phụ thuộc vào số lượng dữ liệu quan sát được cung cấp và mô hình thực tế được huấn luyện. Có 3 cách chia tập dữ liệu thường được sử dụng trên thực tế đó là:

* Hold-out: Đây là phương pháp chỉ chia tập dữ liệu quan sát được cung cấp thành hai tập con, trong đó một tập để huấn luyện (thường chiếm 80% lượng dữ liệu hoặc 70% dữ liệu) và một tập được bỏ ra (Hold-out) ngoài tập huấn luyện đóng vai trò là testing set, là phần còn lại của bộ dữ liệu được cung cấp (thường là 20% lượng dữ liệu hoặc 30% dữ liệu còn lại).
* Train-Validation-Test Split: Đây là phương pháp có sử dụng bộ dữ liệu validation set. Cách thức chia tương tự như trường hợp Hold-out, tức là để riêng ra một bộ dữ liệu để kiểm thử (testing set), và trong lượng dữ liệu còn lại lại chia nhỏ thành 2 phần, 1 phần để train mô hình (thường khoảng 80% lượng dữ liệu còn lại), và một phần để kiểm chứng và hiệu chỉnh mô hình (validation – 20% dữ liệu còn lại).
* Cross Validation: là phương pháp chủ yếu được sử dụng để đánh giá thuật toán huấn luyện hoặc dạng của mô hình huấn luyện chứ không phải mô hình huấn luyện. Khi đó, quá trình huấn luyện và kiểm thử sẽ được thực hiện lặp đi lặp lại nhiều lần. Trong mỗi lần tập dữ liệu quan sát sẽ được lần lượt chia thành 2 phần (training và testing), phần training được sử dụng cho việc huấn luyện, còn phần testing được sử dụng cho kiểm thử.

Trong bài toán nhận diện biển báo giao thông tập dữ liệu không quá lớn, để đảm bảo dữ liệu cho huấn luyện được đầy đủ và cân bằng nên dữ liệu sẽ được phân chia theo tỉ lệ 60/20/20: Trong đó  tập Train và tập Validation chiếm 80% và tập Test chiếm 20% tập dữ liệu.

* Dữ liệu Train lớn nhất vì đây là tập dùng cho huấn luyện mô hình
* Dữ liệu Validation dùng cho đánh giá chất lượng của mô hình, và từ đó có các quyết định thay đổi liên quan đến mô hình (thuật toán học, siêu tham số, đặc trưng, ...)
* Dữ liệu Test dùng cho đánh giá chất lượng của mô hình (nhưng không được dùng để quyết định các thay đổi liên quan đến mô hình - như vậy là gián tiếp tham gia huấn luyện mô hình)
* Kích thước dữ liệu Validation thường bằng dữ liệu test để dữ liệu Validation có thể mô phỏng được dữ liệu test (nếu Validation và tập Test quá khác nhau thì việc hiệu chỉnh mô hình để tốt với Validation không đảm bảo mô hình tốt với Test, như vậy thì Validation sinh ra cũng không có tác dụng)

## **2.4. 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.

## **2.5. Huấn luyện mô hình**

Sau khi đã có dữ liệu tiếp tục sử dụng Transformers để thiết lập quá trình huấn luyện mô hình. Các tham số huấn luyện được thiết lập bao gồm:

* **num\_train\_epochs**: Số vòng lặp (epochs) huấn luyện trên tập dữ liệu. Giá trị được đặt là **4** để đảm bảo quá trình huấn luyện vừa đủ nhanh mà không gây quá tải mô hình.
* **learning\_rate**: Tốc độ học của mô hình, được đặt là **3e-5**, một giá trị phổ biến và phù hợp cho các mô hình BERT-based.
* **weight\_decay**: Tham số để giảm overfitting, với giá trị **0.01**.
* **per\_device\_train\_batch\_size**: Số mẫu trong mỗi batch cho mỗi thiết bị (GPU/CPU) khi huấn luyện, đặt là **4**.
* **gradient\_accumulation\_steps**: Tổng hợp gradient qua 2 bước trước khi cập nhật, tăng hiệu quả khi sử dụng batch nhỏ.

Trong quá trình huấn luyện mô hình, giám sát và điều chỉnh là những bước quan trọng để đảm bảo mô hình hoạt động hiệu quả, tránh các vấn đề như quá khớp, không hội tụ hoặc lãng phí tài nguyên. Trước tiên, cần theo dõi các chỉ số hiệu suất như hàm mất mát (loss) trên tập huấn luyện và tập kiểm tra, cũng như các chỉ số đánh giá cụ thể như Accuracy, F1-score. Nếu loss trên tập huấn luyện giảm nhưng tăng trên tập kiểm tra, đây là dấu hiệu của quá khớp, đòi hỏi các kỹ thuật như Dropout, Weight Decay hoặc tăng cường dữ liệu (Data Augmentation). Ngược lại, nếu cả hai giá trị loss đều cao, mô hình có thể đang underfit, cần được tăng độ phức tạp hoặc cải thiện dữ liệu.

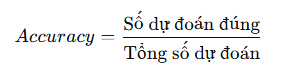
Điều chỉnh siêu tham số như learning rate, batch size, số epoch cũng là yếu tố quan trọng. Learning rate quá lớn khiến mô hình dao động và không hội tụ, trong khi learning rate quá nhỏ làm chậm quá trình học. Sử dụng các kỹ thuật như Early Stopping để dừng huấn luyện khi hiệu suất trên tập kiểm tra không cải thiện sau một số epoch nhất định giúp tiết kiệm thời gian và tránh overfitting. Đồng thời, việc kiểm tra gradient trong quá trình lan truyền ngược giúp phát hiện và xử lý các vấn đề như vanishing gradient hoặc exploding gradient bằng cách áp dụng các kỹ thuật như gradient clipping.

Ngoài ra, cần giám sát thời gian huấn luyện và tài nguyên, đảm bảo batch size phù hợp với bộ nhớ GPU. Các công cụ như TensorBoard hoặc Weights & Biases hỗ trợ trực quan hóa quá trình huấn luyện và quản lý thí nghiệm. Cuối cùng, kiểm tra và làm sạch dữ liệu, vì dữ liệu không đầy đủ hoặc chứa lỗi có thể làm giảm hiệu suất mô hình. Bằng cách kết hợp giám sát chặt chẽ với điều chỉnh hợp lý, mô hình có thể hội tụ nhanh hơn, đạt hiệu suất cao và khả năng tổng quát hóa tốt.

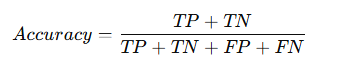
## **2.6. Đánh giá kết quả**

**ACCURACY**

Accuracy là một thuật ngữ được sử dụng trong ngữ cảnh của máy học và học máy để mô tả mức độ đúng đắn của mô hình trong việc dự đoán hay phân loại dữ liệu. Chính xác thường được tính bằng tỷ lệ giữa số lượng dự đoán đúng và tổng số lượng dự đoán. Công thức tính chính xác như sau:



Trong bài toán phân loại nhị phân, công thức này có thể biểu diễn rõ hơn bằng cách sử dụng các giá trị từ ma trận nhầm lẫn (Confusion Matrix):



Điều này đo lường khả năng của mô hình trong việc đưa ra dự đoán chính xác. Accuracy là một phép đo phổ quát và dễ hiểu, nhưng có nhược điểm khi mô hình phải đối mặt với dữ liệu mất cân bằng (imbalance), nơi mà một số lớp xuất hiện nhiều hơn so với các lớp khác. Trong trường hợp này, một mô hình có thể có accuracy cao mà vẫn không thực sự hiệu quả.

**LOSS**

Loss (hàm mất mát) là một khái niệm cốt lõi trong học máy và deep learning, đại diện cho sự khác biệt giữa giá trị dự đoán của mô hình và giá trị thực tế. Mục tiêu của việc tối ưu hóa mô hình là giảm giá trị của hàm mất mát này để cải thiện hiệu suất dự đoán.



Các thuật ngữ khác như "Cost" hoặc "Error" cũng thường được sử dụng thay thế cho "loss". Loss Function (hàm mất mát) chịu trách nhiệm tính toán giá trị mất mát dựa trên dự đoán của mô hình và giá trị thực tế. Cụ thể, loss function đo lường sự khác biệt giữa giá trị dự đoán và giá trị thực tế của mỗi mẫu dữ liệu trong tập huấn luyện.

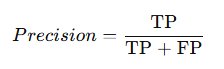
Mục tiêu của quá trình huấn luyện mô hình là tối ưu hóa hàm mất mát tức là giảm thiểu giá trị loss. Quá trình này thường được thực hiện thông qua các thuật toán tối ưu hóa Gradient Descent hoặc các biến thể của nó.

Một số hàm mất mát phổ biến bao gồm:

* Mean Squared Error (MSE): Đo lường trung bình bình phương của sự chênh lệch giữa dự đoán và giá trị thực tế.
* Cross-Entropy Loss (Log Loss): Phổ biến trong các bài toán phân loại, đo lường sự chênh lệch giữa phân phối xác suất dự đoán và phân phối xác suất thực tế.
* Hinge Loss: Thường được sử dụng trong các bài toán Support Vector Machines (SVM) và các bài toán phân loại nhị phân.

**PRECISION**

Precision là tỷ lệ giữa số sample được tính là True Positive (TP) với tổng số sample được phân loại là Positive (bằng chính TP + FP)

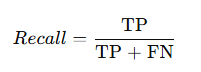


Precision đặc biệt quan trọng trong các bài toán mà FP (False Positive) gây ra hậu quả nghiêm trọng.

Và khi 0< Precision <=1, Precision càng lớn có nghĩa là độ chính xác của các điểm tìm được càng cao.

**RECALL**

Recall là chỉ số đo lường khả năng của mô hình trong việc phát hiện chính xác các mẫu dương tính trong toàn bộ dữ liệu.



Recall quan trọng trong các bài toán mà FN (False Negative) gây ra hậu quả nghiêm trọng.

**F-1 SCORE**

F1-Score là một thước đo kết hợp giữa Precision và Recall. Đặc biệt F1-Score được sử dụng trong trường hợp dữ liệu mất cân bằng, khi số lượng mẫu thuộc các lớp khác nhau không đồng đều. F1-Score đặc biệt hữu ích khi dữ liệu không cân bằng và cần đánh giá hiệu suất tổng quát.



F1-score tự nhiên có giá trị trong khoảng từ 0 đến 1, với 1 là giá trị tốt nhất (đạt được khi cả Precision và Recall đều bằng 1). F1-score giúp cân bằng giữa Precision và Recall, giúp đánh giá hiệu suất của mô hình một cách tổng thể, đặc biệt trong trường hợp có sự mất cân bằng giữa các lớp.

# **CHƯƠNG 3: TRIỂN KHAI VÀ THỰC NGHIỆM**

**3.1. Mô tả bài toán**

## **3.2. Phân tích thiết kế hệ thống**

**3.2.1. Tổng quan chức năng ứng dụng**

**3.2.2. Mô hình phân rã chức năng**

**3.2.3. Giao diện ứng dụng**

## **3.3. Quy trình thực nghiệm**

## **3.4. Đánh giá kết quả**

# **CHƯƠNG 4: KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN**

## **4.1. Kết luận về đề tài**

Trong bối cảnh chuyển đổi số ngày càng mạnh mẽ, việc ứng dụng công nghệ trí tuệ nhân tạo vào lĩnh vực văn bản hành chính không chỉ là một xu hướng mà còn là một nhu cầu cấp thiết. Ứng dụng trả lời văn bản hành chính là một giải pháp sáng tạo, mang lại nhiều lợi ích thiết thực cho cơ quan, tổ chức và người dân.

Sau quá trình tìm hiểu và thực hiện, đồ án đã hoàn thành và giao tiếp và trả lời về văn bản hành chính nói chung và văn bản giao thông hành chính nói riêng. Biết được quá trình thu thập và xử lý dữ liệu để có thể tiến hành huấn luyện trí tuệ nhân tạo. Hoàn thành mục tiêu và kết quả đề ra. Tập dữ liệu vẫn chưa hoàn thiện nên cần phải cập nhật thêm nhiều dữ liệu về các văn bản hành chính khác, đồng thời cũng có thể nâng cấp phần cứng để tăng tốc độ xử lý và hiệu suất của sản phẩm.

**Kết quả đạt được:**

* Hiểu về mô hình Transformer, các dạng kiến trúc trí tuệ nhân tạo như Bert, Bard và áp dụng vào xây dựng mô hình trả lời văn bản hành chính.
* Hiểu cách và các phương pháp thu thập, xử lý dữ liệu trong mô hình.
* Mô hình có khả năng cung cấp thông tin nhanh chóng, chính xác và liên tục , giúp giảm tải công việc cho nhân viên hành chính.
* Người dân và tổ chức có thể tra cứu thông tin hành chính dễ dàng, không cần chờ đợi hoặc thực hiện các thủ tục phức tạp.
* Tiết kiệm về nguồn lực giúp tự động hóa các công việc hành chính cơ bản, giảm thiểu chi phí vận hành và nguồn lực nhân sự.

**Hạn chế:**

* Chương trình hiện tại chưa có khả năng xử lý các câu hỏi phức tạp hoặc các ngữ cảnh đa nghĩa trong văn bản hành chính.
* Vì mang tính nghiên cứu nên hệ thống chỉ mới làm việc trên tập dữ liệu nhỏ với số lượng văn bản có hạn.
* Chưa tích hợp được lên hệ thống hoàn chỉnh để sử dụng trong thực tế tại các cơ quan, tổ chức.

## **4.2. Hướng phát triển của đề tài**

# **TÀI LIỆU THAM KHẢO**

[1] <https://arxiv.org/pdf/1910.13461v1>

[2] John Paul Mueller, Luca Massaron – Artificial Intelligence for Dummies-Wiley (2018)

[3] An Essential Guide to Machine Learning for Beginners By Herbert Jones

[4] https://huggingface.co/

[5] https://github.com/VinAIResearch/BARTpho

[6] https://www.miai.vn/2020/12/01/mi-transformer-tim-hieu-transformer-theo-cach-de-hieu-de-nho/

[7] https://www.miai.vn/2020/12/14/bert-series-chuong-1-bert-la-cai-chi-chi/

# **PHỤ LỤC**

Hướng dẫn sử dụng