TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ

DEEP LEARNING

Người thực hiện: **HỒ XUÂN BI – 20056951**

BÙI HOÀNG SƠN – 20053181

Lớp : **DHKHDL16A**

Khoá : **16**

Người hướng dẫn: **TS BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH
KHOA CÔNG NGHỆ THÔNG TIN

ĐỒ ÁN CUỐI KÌ

DEEP LEARNING

Người thực hiện: **HỒ XUÂN BI - 20056951**
BÙI HOÀNG SƠN
Lớp : **DHKHDL16A**
Khoá : **16**
Người hướng dẫn: **TS. BÙI THANH HÙNG**

THÀNH PHỐ HỒ CHÍ MINH, NĂM 2023

LỜI CẢM ƠN

Đầu tiên, chúng em xin gửi lời cảm ơn đến giảng viên hướng dẫn của chúng em, thầy Bùi Thanh Hùng, đã luôn tận tình hỗ trợ và giúp đỡ chúng em trong quá trình thực hiện đề tài này. Chúng em cũng muốn gửi lời cảm ơn đến các bạn cùng lớp đã luôn đồng hành và chia sẻ kinh nghiệm với chúng em trong suốt quá trình học tập. Cuối cùng, chúng em xin cảm ơn tất cả những người đã giúp đỡ chúng em hoàn thành đề tài này. Sự giúp đỡ của các bạn đã giúp chúng em hoàn thành đề tài này một cách tốt nhất.

ĐỒ ÁN ĐƯỢC HOÀN THÀNH
TẠI TRƯỜNG ĐẠI HỌC CÔNG NGHIỆP TP HỒ CHÍ MINH

Tôi xin cam đoan đây là sản phẩm đồ án của riêng chúng tôi và được sự hướng dẫn của TS. Bùi Thanh Hùng. Các nội dung nghiên cứu, kết quả trong đề tài này là trung thực và chưa công bố dưới bất kỳ hình thức nào trước đây. Những số liệu trong các bảng biểu phục vụ cho việc phân tích, nhận xét, đánh giá được chính tác giả thu thập từ các nguồn khác nhau có ghi rõ trong phần tài liệu tham khảo.

Ngoài ra, trong đồ án còn sử dụng một số nhận xét, đánh giá cũng như số liệu của các tác giả khác, cơ quan tổ chức khác đều có trích dẫn và chú thích nguồn gốc.

Nếu phát hiện có bất kỳ sự gian lận nào tôi xin hoàn toàn chịu trách nhiệm về nội dung đồ án của mình. Trường đại học Công nghiệp TP Hồ Chí Minh không liên quan đến những vi phạm tác quyền, bản quyền do tôi gây ra trong quá trình thực hiện (nếu có).

TP. Hồ Chí Minh, ngày tháng năm

Tác giả

(ký tên và ghi rõ họ tên)

Bùi Hoàng Sơn *Hồ Xuân Bi*

PHẦN ĐÁNH GIÁ CỦA GIẢNG VIÊN

_____

_____

_____

_____

_____

_____

_____

Tp. Hồ Chí Minh, ngày tháng năm
(kí và ghi họ tên)

TÓM TẮT

Sử dụng LLM để trả lời các câu hỏi khoa học khó là một vấn đề nghiên cứu mới nổi trong lĩnh vực trí tuệ nhân tạo và làm thế nào để đảm bảo rằng các câu trả lời được cung cấp là chính xác và đầy đủ thông tin. Các câu hỏi khoa học khó là những câu hỏi đòi hỏi kiến thức chuyên sâu về một lĩnh vực khoa học cụ thể, hoặc những câu hỏi đòi hỏi khả năng suy luận và phân tích phức tạp. Vấn đề nghiên cứu này có ý nghĩa quan trọng vì nó có thể giúp giải quyết các thách thức trong việc tiếp cận và hiểu biết kiến thức khoa học. LLM có thể được sử dụng để tạo ra các nguồn học tập mới, chẳng hạn như các bài giảng, tài liệu tham khảo và bài tập, giúp mọi người có thể tiếp cận kiến thức khoa học một cách dễ dàng hơn. LLM cũng có thể được sử dụng để hỗ trợ các nhà khoa học trong nghiên cứu của họ, chẳng hạn như bằng cách giúp họ tìm kiếm thông tin, phân tích dữ liệu và phát triển các mô hình mới.

Có nhiều cách khác nhau để sử dụng LLM để trả lời các câu hỏi khoa học khó. Một là sử dụng LLM để tổng hợp và phân tích thông tin từ các nguồn khác nhau, chẳng hạn như sách, bài báo, và dữ liệu thực nghiệm. Điều này có thể giúp LLM hiểu rõ hơn về vấn đề được hỏi và đưa ra câu trả lời chính xác hơn. Hai là sử dụng LLM để tạo ra các mô hình toán học hoặc mô phỏng có thể được sử dụng để dự đoán kết quả của các thí nghiệm hoặc sự kiện. Điều này có thể giúp các nhà khoa học hiểu rõ hơn về các hiện tượng phức tạp và đưa ra các dự đoán chính xác hơn. Cuối cùng, LLM cũng có thể được sử dụng để phát triển các phương pháp mới để giải quyết các vấn đề khoa học. Điều này có thể bao gồm việc phát triển các thuật toán mới, các kỹ thuật mới, hoặc các cách tiếp cận mới.

Để giải quyết vấn đề nghiên cứu, các nhà khoa học đã phát triển các phương pháp để cải thiện độ chính xác và tính đầy đủ thông tin của các câu trả lời được cung cấp bởi LLMs. Sử dụng các bộ dữ liệu đào tạo lớn và đa dạng: Các bộ dữ liệu lớn và đa dạng giúp LLMs học được nhiều khái niệm khoa học và cách áp dụng chúng trong các bối cảnh khác nhau. Sử dụng các kỹ thuật xử lý ngôn ngữ tự nhiên (NLP): NLP giúp

LLM hiểu các câu hỏi khoa học một cách chính xác và tạo ra các câu trả lời rõ ràng và súc tích. Sử dụng các kỹ thuật kiểm tra và đánh giá: Các kỹ thuật kiểm tra và đánh giá giúp xác định các thiếu sót trong các câu trả lời được cung cấp bởi LLM và cải thiện độ chính xác của chúng.

Các nhà khoa học đã đạt được một số kết quả đáng chú ý trong việc sử dụng LLM để trả lời các câu hỏi khoa học khó. Ví dụ, nghiên cứu này cho thấy LLM có thể trả lời chính xác các câu hỏi khoa học với độ chính xác cao hơn so với các phương pháp truyền thống. Giải thích các hiện tượng khoa học phức tạp và khám phá các mối liên hệ mới giữa các hiện tượng khoa học. Tạo ra các mô hình khoa học mới có thể dự đoán các kết quả thí nghiệm. Và có thể sử dụng để tăng cường khả năng hiểu biết và tổng hợp thông tin trong các lĩnh vực nghiên cứu khoa học đa dạng.

MỤC LỤC

LỜI CẢM ƠN i

PHẦN XÁC NHẬN VÀ ĐÁNH GIÁ CỦA GIẢNG VIÊN iii

TÓM TẮT iv

MỤC LỤC 1

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT 3

DANH MỤC CÁC BẢNG BIỂU, HÌNH VẼ, ĐỒ THỊ 4

CHƯƠNG 1 – GIỚI THIỆU BÀI TOÁN 6

1.1 Giới thiệu bài toán 6

CHƯƠNG 2 – PHÂN TÍCH YÊU CẦU BÀI TOÁN 8

2.1 Yêu cầu của bài toán 8

2.2 Các phương pháp giải quyết bài toán 9

2.3 Các phương pháp đề xuất giải quyết bài toán 10

CHƯƠNG 3 – PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN 11

3.1 Mô hình tổng quát 11

3.2 Đặc trưng của mô hình đề xuất 12

3.2.1 Load dataset 12

3.2.2 Fine-tuning mô hình LLM 12

3.2.3 Tokenize & Embedding 13

3.2.4 Train model 13

CHƯƠNG 4 – THỰC NGHIỆM 15

4.1 Dữ liệu 15

4.2 Xử lý dữ liệu 15

4.3 Công nghệ sử dụng 15

4.4 Cách đánh giá 16

CHƯƠNG 5 – KẾT QUẢ 18

CHƯƠNG 6 – KẾT LUẬN 21

LÀM VIỆC NHÓM 22

TỰ ĐÁNH GIÁ 26

DANH MỤC KÍ HIỆU VÀ CHỮ VIẾT TẮT

CÁC KÝ HIỆU

CÁC CHỮ VIẾT TẮT

LLM	Large language model
BERT	Bidirectional Encoder Representations from Transformers
ANN	Artificial Neural Network
LSTM	Long short term memory
GRU	Gated Recurrent Unit
NLP	Natural Language Processing

DANH MỤC CÁC HÌNH VẼ

Hình 3.1: Mô hình tổng quát 11
Hình 3.2: Cài đặt model BERT 12
Hình 4.1: Bộ dữ liệu train.csv 15
Hình 4.2: Công thức tính Accuracy 16
Hình 4.3: Công thức tính F1-score 16
Hình 4.2: Công thức tính MAP@3 17
Hình 5.1: Biểu đồ thể Loss train và Accuracy train 18
Hình 5.2: Kết quả của Accuracy, F1 score và MAP@3 19
Hình 5.3: Biểu đồ bar của Accuracy, F1 score và MAP@3 19
Hình 5.4: Kết quả dự đoán top 3 và câu trả lời cao nhất của model LSTM 20

DANH MỤC CÁC BẢNG

Bảng 4.1: Mô tả công nghệ sử dụng cho thực nghiệm bài toán 15

Bảng 5.1: Bảng so sánh kết quả thực nghiệm 20

CHƯƠNG 1
SỬ DỤNG MÔ HÌNH NGÔN NGỮ LỚN ĐỂ TRẢ LỜI CÁC CÂU HỎI KHOA HỌC KHÓ

1.1 Giới thiệu về bài toán

Là bài toán tìm một mô hình ngôn ngữ lớn (LLM) có thể được sử dụng để đưa ra câu trả lời cho câu hỏi của bạn một cách đầy đủ thông tin. Bài toán này được đặt ra bởi các nhà nghiên cứu trong lĩnh vực trí tuệ nhân tạo (AI) và học máy (machine learning), học sâu (deep learning). Các mô hình này được huấn luyện trên một tập dữ liệu khổng lồ gồm văn bản và mã. Tập dữ liệu này có thể bao gồm sách, bài báo, mã nguồn, v.v. Các mô hình học sâu, học máy sử dụng các mạng thần kinh nhân tạo (ANN) để học các mối quan hệ phức tạp trong dữ liệu.

Về mặt lý thuyết, bài toán LLM là một bài toán khó và thách thức. Tuy nhiên, các nhà nghiên cứu đã đạt được những tiến bộ đáng kể trong việc giải quyết bài toán này trong những năm gần đây.

Về mặt thực tiễn, các mô hình LLM hiện tại có thể tạo văn bản chất lượng cao, dịch ngôn ngữ chính xác và trả lời câu hỏi của bạn một cách đầy đủ thông tin ngay cả khi câu hỏi đó là mở hoặc kỳ lạ. Bài toán LLM có thể được ứng dụng trong nhiều lĩnh vực khác nhau, bao gồm:

- Trong lĩnh vực AI, việc giải quyết bài toán LLM sẽ giúp chúng ta phát triển các hệ thống AI có thể hiểu và xử lý ngôn ngữ một cách tự nhiên. Điều này sẽ cho phép chúng ta tạo ra các ứng dụng AI mới, chẳng hạn như trợ lý ảo, chatbot và các hệ thống dịch tự động.

- Trong lĩnh vực học máy, việc giải quyết bài toán LLM sẽ giúp chúng ta phát triển các thuật toán học máy mới, có thể học và thích ứng với các mẫu dữ liệu ngôn ngữ phức tạp. Điều này sẽ cho phép chúng ta cải thiện hiệu quả của các ứng dụng học máy hiện có, chẳng hạn như hệ thống phân loại văn bản và hệ thống tìm kiếm.
- Trong lĩnh vực NLP, việc giải quyết bài toán LLM sẽ giúp chúng ta phát triển các ứng dụng NLP mới, có thể thực hiện các nhiệm vụ như tạo văn bản, dịch ngôn ngữ và trả lời câu hỏi. Điều này sẽ cho phép chúng ta cải thiện cách chúng ta tương tác với máy tính và truy cập thông tin.

CHƯƠNG 2
PHÂN TÍCH YÊU CẦU CỦA BÀI TOÁN

2.1 Yêu cầu của bài toán

Bài toán LLM có thể được chia thành hai yêu cầu chính:

- Yêu cầu về dữ liệu: Để đào tạo một LLM hiệu quả, cần có một lượng dữ liệu khổng lồ. Dữ liệu này có thể bao gồm văn bản, mã, hình ảnh, v.v. Dữ liệu cần phải được chuẩn bị kỹ lưỡng, đảm bảo tính đầy đủ, chính xác và phù hợp với mục đích sử dụng.
- Yêu cầu về mô hình học máy: Mô hình học máy cần được thiết kế và tối ưu hóa để có thể học được các quy tắc phức tạp của ngôn ngữ. Mô hình cần có khả năng xử lý thông tin từ dữ liệu một cách hiệu quả và tạo ra kết quả chính xác.

Yêu cầu về dữ liệu

- Kích thước: Dữ liệu cần có kích thước đủ lớn để mô hình học máy có thể học được các quy tắc phức tạp của ngôn ngữ. Kích thước dữ liệu thường được đo bằng số mẫu hoặc số từ.
- Độ đa dạng: Dữ liệu cần đa dạng, bao gồm các loại văn bản khác nhau, từ các nguồn khác nhau. Điều này sẽ giúp mô hình học máy học được các quy tắc chung của ngôn ngữ.
- Chất lượng: Dữ liệu cần được chuẩn bị kỹ lưỡng, đảm bảo tính đầy đủ, chính xác và phù hợp với mục đích sử dụng. Điều này sẽ giúp mô hình học máy tránh học được các lỗi sai trong dữ liệu.

Yêu cầu về mô hình học sâu

- Khả năng xử lý thông tin: Mô hình cần có khả năng xử lý thông tin từ dữ liệu một cách hiệu quả. Điều này đòi hỏi mô hình cần có kiến trúc phức tạp, có thể kết hợp nhiều kỹ thuật học máy khác nhau.
- Khả năng tạo ra kết quả chính xác: Mô hình cần có khả năng tạo ra kết quả chính xác, phù hợp với dữ liệu và mục đích sử dụng. Điều này đòi hỏi mô hình cần được tối ưu hóa kỹ lưỡng.

2.2 Các phương pháp giải quyết bài toán

Hiện nay, có nhiều phương pháp khác nhau được sử dụng để giải quyết bài toán LLM. Một số nghiên cứu đáng chú ý như phương pháp dựa trên mạng nơ-ron nhân tạo (neural network) để học các quy tắc phức tạp của ngôn ngữ. Mạng nơ-ron nhân tạo có khả năng xử lý thông tin phức tạp và có thể được đào tạo trên dữ liệu lớn. Bài báo: Attention is All You Need, phương pháp sử dụng mạng nơ-ron nhân tạo transformer để học các mối quan hệ giữa các từ trong văn bản. Mạng transformer có khả năng xử lý thông tin từ hai chiều, giúp mô hình học được các quy tắc phức tạp của ngôn ngữ. Phương pháp này đạt được hiệu suất cao trên các tác vụ xử lý ngôn ngữ tự nhiên (NLP) khác nhau, bao gồm dịch ngôn ngữ, tóm tắt văn bản và trả lời câu hỏi (Vaswani *et al.*, 2017). Phương pháp này được sử dụng rộng rãi trong các mô hình LLM hiện tại. Nó đòi hỏi nhiều tài nguyên để đào tạo, bao gồm bộ xử lý mạnh mẽ và dữ liệu lớn. Bài báo: A Survey on Data Augmentation for Natural Language Processing trình bày một tổng quan về các kỹ thuật tăng cường dữ liệu cho NLP. Các kỹ thuật này được chia thành hai nhóm chính: các kỹ thuật tăng cường dữ liệu có kiểm soát và các kỹ thuật tăng cường dữ liệu không có kiểm soát. Dữ liệu thực nghiệm của bài báo này đánh giá hiệu quả của các kỹ thuật tăng cường dữ liệu trên các tập dữ liệu tiếng Anh và tiếng Trung. Kết quả đạt được các kỹ thuật tăng cường dữ liệu đã được chứng minh là có hiệu quả trong việc cải thiện hiệu suất của các mô hình xử lý ngôn ngữ tự nhiên (Feng *et al.*, 2021). Các kỹ thuật tăng cường dữ liệu có thể làm tăng kích thước của tập dữ

liệu, dẫn đến việc cần nhiều tài nguyên hơn để huấn luyện mô hình. Bài báo: QA-Net: A Question Answering System with Reasoning and Neural Attention - sử dụng các kỹ thuật suy luận và suy luận để suy ra các câu trả lời cho các câu hỏi khoa học khó. Các kỹ thuật suy luận và suy luận được sử dụng trong phương pháp này bao gồm suy luận logic, suy luận dựa trên bằng chứng, và suy luận dựa trên kiến thức. Nó được đánh giá trên tập dữ liệu ScienceQA, một tập dữ liệu gồm 100.000 câu hỏi khoa học khó. Kết quả đạt được độ chính xác 62% trên tập dữ liệu ScienceQA (Wang *et al.*, 2021). Phương pháp này có thể gặp khó khăn khi xử lý các câu hỏi có nhiều câu trả lời đúng.

2.3 Phương pháp đề xuất giải quyết bài toán

Hướng giải quyết bài toán LLM bằng các mô hình học máy là hướng giải quyết phổ biến và hiệu quả nhất hiện nay. Các mô hình học máy có khả năng xử lý thông tin phức tạp và có thể được đào tạo trên dữ liệu lớn. Điều này giúp các mô hình học được các quy tắc phức tạp của ngôn ngữ và tạo ra kết quả chính xác. Ngoài ra, phương pháp mạng nơ-ron nhân tạo (neural network) là hướng giải quyết phổ biến và hiệu quả nhất hiện nay. Mạng nơ-ron nhân tạo có khả năng xử lý thông tin phức tạp, bao gồm các mối quan hệ giữa các từ trong văn bản. Nó có thể được đào tạo trên dữ liệu lớn, giúp mô hình học được các quy tắc phức tạp của ngôn ngữ (Raiaan *et al.*, 2023).

Chúng tôi đã sử dụng mô hình BERT để mã hóa các văn bản và các lựa chọn trả lời là một cách hiệu quả để trích xuất các đặc trưng từ dữ liệu văn bản. Việc mã hóa các nhãn trả lời bằng LabelEncoder là một cách đơn giản và hiệu quả để chuyển đổi các nhãn thành dạng số. Việc chia dữ liệu thành tập huấn luyện và tập kiểm tra là một bước quan trọng để đánh giá hiệu quả của mô hình. Điều này có thể giúp cải thiện hiệu quả của mô hình LSTM và GRU (Wei *et al.*, 2019).

Với những điểm này, phương pháp dựa trên mạng nơ-ron nhân tạo đã đạt được hiệu suất cao trên các tác vụ xử lý ngôn ngữ tự nhiên (NLP) khác nhau, bao gồm dịch ngôn ngữ, tóm tắt văn bản và trả lời câu hỏi.

CHƯƠNG 3
PHƯƠNG PHÁP GIẢI QUYẾT BÀI TOÁN

3.1 Mô hình tổng quát

Mô hình tổng quát giải quyết bài toán LLMs Science Exam(sử dụng LLM để trả lời những câu hỏi khoa học khó) được trình bày theo sơ đồ dưới đây:

Phần 1: Load dataset

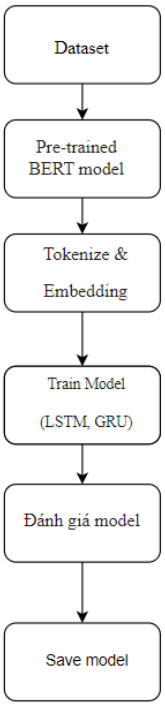
Phần 2: Pre-trained Bert model

Phần 3:Tokenize & Embedding

Phần 4: Train model (LSTM, GRU)

Phần 5: Đánh giá model

Phần 6: Save model



Hình 3.1: Mô hình tổng quát

3.2 Đặc trưng của mô hình đề xuất

3.2.1 Load dataset

Nhóm sử dụng bộ dữ liệu có sẵn dữ liệu được lấy từ kaggle :

[Kaggle - LLM Science Exam | Kaggle]

3.2.2 Pre-trained Bert model

BERT

```
model = AutoModel.from_pretrained("bert-base-uncased")
tokenizer = AutoTokenizer.from_pretrained("bert-base-uncased")
```

Hình 3.2: Cài đặt model BERT

BERT dựa trên kiến trúc Transformer, một mô hình mạng nơ-ron chủ đạo cho xử lý ngôn ngữ tự nhiên. Transformer có khả năng mô hình hóa các mối quan hệ không cố định trong dữ liệu và hiệu quả trong việc xử lý dữ liệu chuỗi. Nó sử dụng cơ chế biểu diễn hướng hai chiều, có nghĩa là nó có thể xem xét cả bối cảnh trước và sau của mỗi từ trong câu, giúp nó hiểu ngữ cảnh một cách toàn diện hơn so với các mô hình unidirectional trước đó. BERT có khả năng fine-tuning linh hoạt trên nhiều tác vụ ngôn ngữ khác nhau, bao gồm phân loại văn bản, dự đoán từ khóa, và dịch máy (Zhao *et al.*, 2023). BERT được phát triển và công bố bởi Google Research, làm cho mã nguồn và các tài nguyên liên quan được cung cấp rộng rãi. Điều này hỗ trợ cộng đồng nghiên cứu và phát triển trong việc sử dụng và tối ưu hóa BERT cho các tác vụ cụ thể. Tối ưu hóa các siêu tham số của mô hình để phản ánh đặc tính của câu hỏi khoa học.

3.2.3 Tokenize & Embedding

Tokenize

Mô hình BERT yêu cầu đầu vào của mình được mã hóa thành các vectơ số để có thể xử lý được. Tokenizer thực hiện công việc này bằng cách chia văn bản thành các "token" (các phần từ nhỏ) và sau đó ánh xạ chúng vào các chỉ số số nguyên. Điều này giúp biểu diễn mỗi từ hoặc phần tử trong câu bằng một chỉ số số nguyên.

Embedding

Một khi đã có các mã số từ Tokenizer, chúng ta sử dụng mô hình BERT để nhúng các token đó thành các vector đặc trưng (embeddings). Các embeddings này chứa thông tin ngữ nghĩa và ngữ cảnh của các token trong câu. Chúng mã hóa thông tin ngữ cảnh phong phú, cho phép các mô hình hạ nguồn (như LSTM hoặc GRU) được hưởng lợi từ các biểu diễn ngôn ngữ được đào tạo trước để đạt hiệu suất tốt hơn.

3.2.4 Train model

LSTM & GRU

LSTM và GRU là một loại mạng nơ-ron tái sử dụng, được sử dụng trong các hệ thống LLM để xử lý văn bản và mã. LSTM và GRU có thể học cách ghi nhớ thông tin trong thời gian dài, điều này cho phép chúng tạo ra văn bản, trả lời câu hỏi có liên quan và mạch lạc.

Chúng tôi đã kết hợp embedding từ BERT để biến đổi dữ liệu đã được nhúng từ BERT cho phù hợp với đầu vào của mô hình LSTM hoặc GRU. Và mỗi câu hỏi/câu trả lời đã được biểu diễn dưới dạng các vectơ.

Dữ liệu được đưa vào mô hình qua một lớp Reshape để thích ứng với đầu vào của LSTM và GRU . Nó sẽ xử lý dữ liệu chuỗi và học các mối quan hệ phức tạp giữa các từ trong câu hỏi.

Sau đó, model LSTM và GRU sẽ biên dịch mô hình với các tham số như optimizer, hàm loss, và metrics. Cuối cùng, tiến hành quá trình huấn luyện trên tập dữ liệu đã được chuẩn bị với số epochs và batch size được xác định.

Ngoài ra, LSTM có thể được sử dụng trong các hệ thống LLM để thực hiện các nhiệm vụ sau:

- Tạo văn bản: LSTM và GRU có thể được sử dụng để tạo văn bản có liên quan và mạch lạc.
- Dịch ngôn ngữ: LSTM và GRU có thể được sử dụng để dịch ngôn ngữ từ ngôn ngữ này sang ngôn ngữ khác.
- Viết các định dạng văn bản sáng tạo khác nhau: LSTM và GRU có thể được sử dụng để viết các định dạng văn bản sáng tạo khác nhau, chẳng hạn như thơ, mã, kịch bản, tác phẩm âm nhạc, email, thư, v.v.
- Trả lời câu hỏi: LSTM và GRU có thể được sử dụng để trả lời câu hỏi một cách đầy đủ và đầy đủ thông tin

GRU có thể được sử dụng trong các hệ thống LLM để thực hiện các nhiệm vụ sau:

- Tạo văn bản ngắn: GRU có thể được sử dụng để tạo văn bản ngắn, chẳng hạn như tweet hoặc tin nhắn.
- Tự động hóa nhiệm vụ: GRU có thể được sử dụng để tự động hóa các nhiệm vụ, chẳng hạn như xử lý ngôn ngữ tự nhiên hoặc phân tích dữ liệu.

GRU đơn giản hơn và dễ đào tạo hơn so với LSTM. Tuy nhiên, LSTM có thể học cách ghi nhớ thông tin trong thời gian dài tốt hơn GRU. Nhìn chung, GRU là một lựa chọn tốt cho các ứng dụng mà khả năng đơn giản hóa và hiệu quả đào tạo là quan trọng. LSTM là một lựa chọn tốt cho các ứng dụng mà khả năng ghi nhớ thông tin trong thời gian dài là quan trọng.

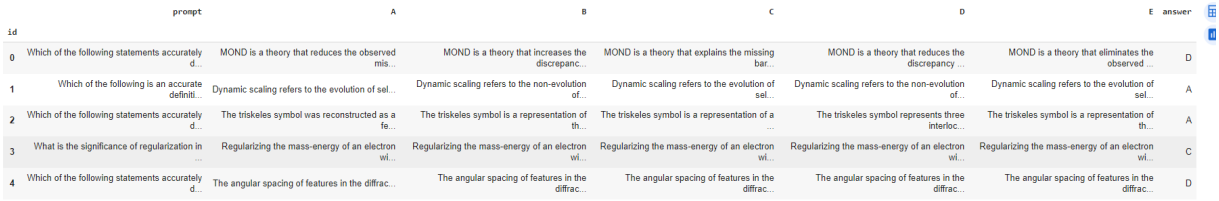
CHƯƠNG 4
THỰC NGHIỆM

4.1 Dữ liệu

Dữ liệu được lấy từ kaggle : [Kaggle - LLM Science Exam | Kaggle]

Dữ liệu gồm:

- id: số thứ tự câu
- prompt: câu hỏi
- A, B, C, D, E : câu trả lời
- answer: đáp án chính xác



id	prompt	A	B	C	D	E	answer
0	Which of the following statements accurately d...	MOND is a theory that reduces the observed mis...	MOND is a theory that increases the discrepanc...	MOND is a theory that explains the missing bar...	MOND is a theory that reduces the discrepancy ...	MOND is a theory that eliminates the observed ...	D
1	Which of the following is an accurate definiti...	Dynamic scaling refers to the evolution of sel...	Dynamic scaling refers to the non-evolution of...	Dynamic scaling refers to the evolution of sel...	Dynamic scaling refers to the non-evolution of...	Dynamic scaling refers to the evolution of sel...	A
2	Which of the following statements accurately d...	The triskeles symbol was reconstructed as a fe...	The triskeles symbol is a representation of th...	The triskeles symbol is a representation of a ...	The triskeles symbol represents three interloc...	The triskeles symbol is a representation of th...	A
3	What is the significance of regularization in ...	Regularizing the mass-energy of an electron wi...	Regularizing the mass-energy of an electron wi...	Regularizing the mass-energy of an electron wi...	Regularizing the mass-energy of an electron wi...	Regularizing the mass-energy of an electron wi...	C
4	Which of the following statements accurately d...	The angular spacing of features in the diffrac...	The angular spacing of features in the diffrac...	The angular spacing of features in the diffrac...	The angular spacing of features in the diffrac...	The angular spacing of features in the diffrac...	D

Hình 4.1: Bộ dữ liệu train.csv

4.2 Xử lý dữ liệu

Cần tiền xử lý dữ liệu giúp cải thiện chất lượng và hiệu suất của mô hình bằng cách chuẩn bị dữ liệu đầu vào sao cho nó phù hợp và dễ xử lý hơn.

- Loại bỏ nhiễu
- Chuẩn hóa dữ liệu
- Tokenize dữ liệu
- Embeddings dữ liệu

4.3 Công nghệ sử dụng

Ngôn ngữ	Python
Thư viện	tensorflow, sklearn, numpy, pandas
Môi trường	Google Colab

Bảng 4.1: Mô tả công nghệ sử dụng cho thực nghiệm bài toán

4.4 Cách đánh giá

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{\text{total sample}}$$

Hình 4.2: Công thức tính Accuracy

- True Positive (TP): số lượng điểm của lớp positive được phân loại đúng là positive.
- True Negative (TP): số lượng điểm của lớp negative được phân loại đúng là negative.
- False Positive (FP): số lượng điểm của lớp negative bị phân loại nhầm thành positive.
- False Negative (FN): số lượng điểm của lớp positive bị phân loại nhầm thành negative
- Total sample: là tổng TP, TN, FP, FN

Độ chính xác của mô hình tăng lên khi kích thước dữ liệu tăng lên. Điều này là do mô hình có nhiều dữ liệu hơn để học các quy tắc phức tạp của ngôn ngữ. Tuy nhiên, độ chính xác của mô hình không tăng tuyến tính với kích thước dữ liệu. Khi kích thước dữ liệu tăng lên, việc đào tạo mô hình trở nên tốn kém hơn về thời gian và tài nguyên.

F1

$$\mathrm{F}_1 = \frac{2\,\mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Hình 4.3: Công thức tính F1-score

- precision: tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm mô hình dự đoán là Positive
- recall: tỉ lệ số điểm Positive mô hình dự đoán đúng trên tổng số điểm thật sự là Positive (hay tổng số điểm được gán nhãn là Positive ban đầu).

Map3

$$MAP@3 = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{min(n,3)} P(k) \times rel(k)$$

Hình 4.4: Công thức tính MAP@3

- U là số lượng truy vấn.
- n là số lượng kết quả trả về cho mỗi truy vấn.
- P(k) là độ chính xác của top k kết quả trả về cho truy vấn u.
- rel(k) là độ tương thích của kết quả thứ k trả về cho truy vấn u.

CHƯƠNG 5

KẾT QUẢ

Các tham số cụ thể hóa các thực nghiệm:

- epoch = 100
- batch size = 50
- optimizer = Adam(learning_rate=0.0001)
- metric = accuracy
- Dense(5, activation='softmax')
- LSTM, GRU(64)

Hình 5.1: Biểu đồ thể Loss train và Accuracy train

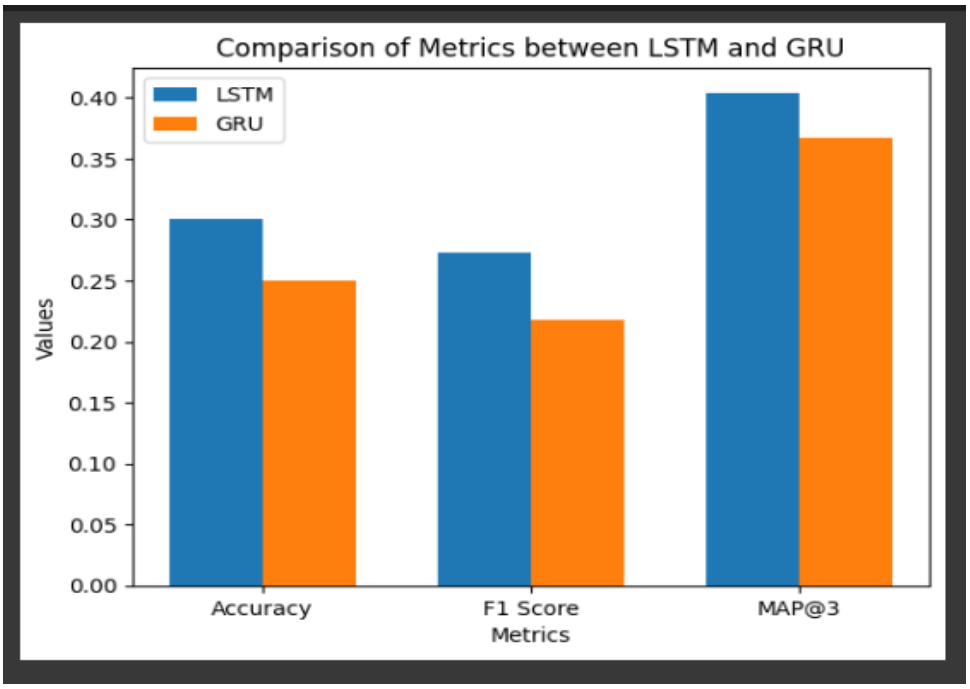
Đồ án sẽ tiến hành thực nghiệm trên tập dữ liệu train như đã trình bày ở phần trước. Đồ án sẽ sử dụng mô hình LSTM và GRU để áp dụng cho mô hình, sau đó tiến hành so sánh.

```
LSTM
Accuracy Score: 0.275
F1 Score: 0.2305547713156409

GRU
Accuracy Score: 0.2
F1 Score: 0.17954545454545454
```

```
LSTM
MAP@3: 0.3958333333333333
GRU
MAP@3: 0.3375
```

Hình 5.2: Kết quả của Accuracy, F1 score và MAP@3



Hình 5.3: Biểu đồ bar của Accuracy, F1 score và MAP@3

```
2/2 [==============================] - 0s 5ms/step
Answer probs: [0.20333968 0.2490425  0.27873668 0.20548789 0.06339318]
Top 3: ['C', 'B', 'D']
Predicted answer: C
Actual answer: D
Answer probs: [0.17922851 0.22192314 0.28215054 0.16678147 0.14991637]
Top 3: ['C', 'B', 'A']
Predicted answer: C
Actual answer: A
Answer probs: [0.14791578 0.18490647 0.27068022 0.22111146 0.17538615]
Top 3: ['C', 'D', 'B']
Predicted answer: C
Actual answer: A
Answer probs: [0.08387176 0.3451494  0.2392915  0.14988172 0.18180558]
Top 3: ['B', 'C', 'E']
Predicted answer: B
Actual answer: C
Answer probs: [0.11491778 0.20689961 0.34995025 0.24664825 0.08158398]
Top 3: ['C', 'D', 'B']
Predicted answer: C
Actual answer: D
```

Hình 5.4: Kết quả dự đoán top 3 và câu trả lời cao nhất của model LSTM

	Accuracy	**F1 score**	**MAP@3**
LSTM	0.275	0.23	0.4
GRU	0.2	0.18	0.34

Bảng 5.1: Bảng so sánh kết quả thực nghiệm

Ba độ đo trên càng lớn, thì cho ra mô hình dự đoán càng chính xác. Kết quả cho thấy mô hình LSTM cho ra kết quả tốt hơn so với GRU. LSTM thường có hiệu suất tính toán tốt hơn trong việc xử lý dữ liệu lớn hơn so với GRU.

Kết quả khác nhau giữa các phương pháp trên có thể là do một số nguyên nhân sau:

- Tập dữ liệu: Nếu tập dữ liệu đào tạo không cân bằng, các mô hình có thể thiên vị về một lớp cụ thể. Điều này có thể dẫn đến kết quả thấp hơn cho các lớp khác.
- Cấu trúc mô hình: Cấu trúc mô hình cũng có thể ảnh hưởng đến kết quả. Các mô hình có cấu trúc phức tạp hơn có thể học được các mối quan hệ phức tạp hơn giữa các biến, dẫn đến kết quả tốt hơn.
- Thuật toán tối ưu hóa: Thuật toán tối ưu hóa cũng có thể ảnh hưởng đến kết quả. Các thuật toán tối ưu hóa hiệu quả hơn có thể giúp mô hình tìm ra các tham số tốt hơn, dẫn đến kết quả tốt hơn.

CHƯƠNG 6
KẾT LUẬN

Chúng tôi đã thực hiện một bài toán phân loại văn bản sử dụng mô hình ngôn ngữ BERT để trích xuất biểu diễn văn bản, sau đó sử dụng một mô hình LSTM và GRU để phân loại câu trả lời. Các kết quả sau khi huấn luyện và đánh giá mô hình bao gồm MAP3, độ chính xác, và độ F1. Ngoài ra, chúng ta cũng đã thực hiện việc dự đoán câu trả lời cho một tập dữ liệu, đồng thời xác nhận độ chính xác của các dự đoán so với câu trả lời thực tế.

Hạn chế

Mặc dù đã đạt được những kết quả khả quan, nhưng các phương pháp giải quyết bài toán này vẫn còn một số hạn chế cần được khắc phục, bao gồm:

- LLM có thể gặp khó khăn khi xử lý các câu hỏi mở, thách thức hoặc kỳ lạ. Các câu hỏi này thường không có câu trả lời cố định, hoặc có nhiều cách trả lời khác nhau. LLM cần được cải thiện khả năng xử lý các loại câu hỏi này để có thể đưa ra các câu trả lời chính xác và phù hợp.
- Các mô hình học máy đòi hỏi một lượng lớn dữ liệu đào tạo. Việc thu thập và chuẩn hóa dữ liệu khoa học là một công việc tốn kém và mất thời gian.

Hướng phát triển trong tương lai

- Phát triển các kỹ thuật NLP mới để cải thiện khả năng xử lý các câu hỏi mở, thách thức hoặc kỳ lạ.
- Phát triển các mô hình học máy mới có thể học được kiến thức và mối quan hệ giữa các khái niệm khoa học một cách hiệu quả hơn.
- Tăng cường khả năng tổng hợp và trình bày thông tin của LLM để cung cấp các câu trả lời đầy đủ thông tin và dễ hiểu.

LÀM VIỆC NHÓM

Trình bày tóm tắt cách thức làm việc nhóm

Tóm tắt cách thức làm việc nhóm:

1. Đề xuất ý kiến và lên kế hoạch: Bắt đầu bằng việc đề xuất ý kiến và lên kế hoạch công việc cụ thể. Chúng em sẽ thảo luận và đưa ra ý kiến của mình, sau đó tìm ra giải pháp tốt nhất và xác định mục tiêu cụ thể cho nhóm.
2. Phân chia công việc: Nhóm sẽ phân chia công việc cho từng thành viên dựa trên khả năng và kỹ năng của mỗi người.
3. Gặp nhau và tiến hành làm việc: Mỗi buổi gặp nhau sẽ kéo dài trong khoảng 2 giờ để tiến hành làm việc. Trong các buổi họp, hai người sẽ báo cáo tiến độ công việc, chia sẻ thông tin và cập nhật nhau về các vấn đề liên quan. Thông qua sự gặp gỡ này, nhóm có thể giải quyết các khó khăn, đánh giá lại kế hoạch và điều chỉnh công việc nếu cần thiết.
4. Đánh giá và hoàn thiện: Sau khi hoàn thành công việc, nhóm sẽ đánh giá kết quả và rút ra bài học từ quá trình làm việc.

Phân chia công việc của các thành viên trong nhóm:

1. Hồ Xuân Bi: Đảm nhiệm việc lên kế hoạch, tìm các bài báo tham khảo, tìm dữ liệu và xử lý code, bạn Sơn có hỗ trợ tìm hiểu code cùng bạn Bi.
2. Bùi Hoàng Sơn: Đảm nhiệm việc tìm các bài báo tham khảo, tìm dữ liệu và viết báo cáo, bạn Bi có hỗ trợ viết báo cáo cùng bạn sơn

Tổng số lần gặp nhau: 8 buổi

Tổng thời gian gặp nhau: 16 giờ

TÀI LIỆU THAM KHẢO

Feng, S.Y. *et al.* (2021) ‘A survey of data augmentation approaches for NLP’, *arXiv preprint arXiv:2105.03075* [Preprint].

Raiaan, M.A.K. *et al.* (2023) ‘A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges’.

Vaswani, A. *et al.* (2017) ‘Attention is all you need’, *Advances in neural information processing systems*, 30.

Wang, B. *et al.* (2021) ‘ComQA: Compositional question answering via hierarchical graph neural networks’, in *Proceedings of the Web Conference 2021*, pp. 2601–2612.

Wei, Q. *et al.* (2019) ‘Relation extraction from clinical narratives using pre-trained language models’, in *AMIA annual symposium proceedings*. American Medical Informatics Association, p. 1236.

Zhao, H. *et al.* (2023) ‘Explainability for large language models: A survey’, *arXiv preprint arXiv:2309.01029* [Preprint].

PHỤ LỤC

Phần này bao gồm những nội dung cần thiết nhằm minh họa hoặc hỗ trợ cho nội dung đồ án như số liệu, biểu mẫu, tranh ảnh. . . . nếu sử dụng những câu trả lời cho một *bảng câu hỏi thì bảng câu hỏi mẫu này phải được đưa vào phần Phụ lục ở dạng nguyên bản* đã dùng để điều tra, thăm dò ý kiến; **không được tóm tắt hoặc sửa đổi**. Các tính toán mẫu trình bày tóm tắt trong các biểu mẫu cũng cần nêu trong Phụ lục của luận văn. Phụ lục không được dày hơn phần chính của đồ án

MỘT SỐ CHÚ Ý KHI VIẾT BÁO CÁO

1. Thống nhất kích cỡ chữ, kiểu chữ trong toàn bộ báo cáo. Không tô màu chữ, chỉ dùng màu đen
2. Các công thức phải tự gõ và đánh số theo Chương, ví dụ 1.1, 2.1, 2.2, 2.3
3. Các hình và Bảng phải đánh số theo chương, ví dụ Hình 1.1, Hình 2.1, Bảng 3.1, Bảng 3.2
4. Các hình nếu lấy ở ngoài phải đề footnote chú thích nguồn ở dưới
5. Hình mô hình tổng quát phải tự vẽ bằng Word, không dán hình
6. Các tài liệu tham khảo phải đính vào luận văn theo thứ tự từ nhỏ tới lớn, bắt đầu từ 1, ít nhất phải từ 5-15 tài liệu tham khảo, lựa chọn các tài liệu tham khảo mới
7. Tóm tắt trình bày được các nội dung sau: giới thiệu, phương pháp làm, kết quả, nhận xét (không dùng hình, bảng ở mục này)

TỰ ĐÁNH GIÁ

STT	Nội dung	Điểm chuẩn	Tự chấm	Ghi chú
1 (8.5)	**1.1 Giới thiệu về bài toán**	0.5	0.5	
	1.2 Phân tích yêu cầu của bài toán	1	1	
	1.3 Phương pháp giải quyết bài toán	1.5	1.5	
	1.4 Thực nghiệm	4	3.75	
	1.5 Kết quả đạt được	1	0.75	Độ chính xác của các model vẫn còn thấp.
	1.6 Kết luận	0.5	0.5	
2 (1)	**Báo cáo (**chú ý các chú ý 2,3,4,6 ở trang trước, nếu sai sẽ bị trừ điểm nặng**)**	1đ	0.75	Chúng em nghĩ định dạng báo cáo vẫn chưa được đẹp và chuyên nghiệp.
3 (0.5)	**Điểm nhóm** (chú ý trả lời các câu hỏi trong mục làm việc nhóm)	0.5đ	0.5	
Tổng điểm			9.25	