PHÁT TRIỂN HỆ THỐNG Q-A VỀ LUẬT VI PHẠM LIÊN QUAN ĐẾN XE MÁY TẠI VIỆT NAM

Nguyễn Quang Huy - 230201014

Tóm tắt

- Lớp: CS2205.CH181
- Link Github: https://github.com/huynguyen261219/CS2205
- Link YouTube video: https://www.youtube.com/watch?v=nrgrmTeVq8c
- Ảnh + Họ và Tên: Nguyễn Quang Huy

Giới thiệu

- Xe máy là phương tiện giao thông chính tại Việt Nam. Tuy nhiên nhiều người dân có thể vẫn chưa biết hết về luật vi phạm liên quan đến xe máy dẫn đến vô tình bị cảnh sát giao thông xử phạt.
- Vì vậy, cần phải có 1 ứng dụng có thể giúp người dân có thể hỏi đáp về các trường hợp liên quan đến quy định về xe máy khi tham gia giao thông

=> Phát triển 1 con BOT hỏi đáp bằng Tiếng Việt sử dụng PhoBERT

Mục tiêu

- Thu thập dữ liệu text về văn bản luật giao thông, các tài liệu hướng dẫn và các giải thích rõ ràng về các vi phạm giao thông liên quan đến xe máy và lưu trữ trên dịch vụ cloud Google Firestore có hỗ trợ embedding
- Cải tiến mô hình với dữ liệu được thu thập để có độ chính xác cao nhất, độ trễ thấp và lượng thông tin phải trọng tâm và đầy đủ nhất
- Xây dựng ứng dụng thuận tiện cho user hỏi đáp với độ trễ < 5s
- Update lại dataset khi luật thay đổi

Nội dung và Phương pháp

- Tiền xử lý dữ liệu
  - Loại bỏ các thông tin nhiễu (chấm câu, ký tự đặc biệt)
  - Tokenize
  - Thu thập từ nhiều nguồn để phân chia thành tập train, test, val
- Sau đó fine-tune vào mô hình PhoBERT

Nội dung và Phương pháp

- Chọn mô hình PhoBERT-large để xử lý các câu hỏi phức tạp
- Thêm lớp Attention vào cuối của PhoBERT để tập trung vào lấy các phần quan trọng sau khi biểu diễn văn bản và câu hỏi thành vector
- Sau khi fine-tune, thực hiện đánh giá bằng phương pháp EM (Exact Match), F1 Score và test xem performance
- Cải thiện và tối ưu thời gian câu trả lời sinh ra
- Deploy và tích hợp vào trong ứng dụng Android và IOS

Kết quả dự kiến

- Dataset chứa đầy đủ thông tin về luật giao thông liên quan đến xe máy: các văn bản, thông tin trên website chính phủ
- Mô hình trả lời chính xác trên các câu hỏi của user và các biến thể của câu hỏi như: viết tắt, đa nghĩa, ngữ cảnh
- Thời gian trả lời có thể lâu hơn nhưng phải dưới 10s
- Tối ưu kích thước mô hình cho ứng dụng di động

Tài liệu tham khảo

- [1]. Dat Quoc Nguyen and Anh Tuan Nguyen (2020). PhoBERT: Pre-trained language models for Vietnamese. ArXiv preprint arXiv:2003.00744.
- [2]. Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv preprint arXiv:1810.04805
- [3]. Thi-Thanh Ha, Van-Nha Nguyen, Kiem-Hieu Nguyen, Kim-Anh Nguyen and Tien-Thanh Nguyen (2020). Utilizing BERT for Question Retrieval in Vietnamese E-commerce Sites. ACL