Báo cáo Thực hành 06: Xây dựng và Huấn luyện Mô hình Dịch Máy Tiếng Anh - Tiếng Việt với Học Sâu

1. Mục tiêu.

Mục tiêu chính của bài thực hành là xây dựng và huấn luyện các mô hình dịch máy từ tiếng Anh sang tiếng Việt sử dụng học sâu, cụ thể gồm:

- Hiện thực một mô hình RNN cơ bản (Seq2Seq với GRU).
- Hiện thực một mô hình học sâu tiên tiến, ưu tiên các kiến trúc state-of-the-art (SOTA) như Transformer.
- Thử nghiệm ít nhất 5 cấu hình siêu tham số khác nhau cho mỗi mô hình, đánh giá hiệu quả dựa trên các chỉ số loss và BLEU.
- Theo dõi và trực quan hóa quá trình huấn luyện thông qua công cụ wandb.
- Xây dựng giao diện demo trực quan cho phép người dùng nhập câu tiếng Anh và nhận câu dịch tiếng Việt từ mô hình.

2. Phương pháp và triển khai.

2.1. Tiền xử lý dữ liệu.

- Tập dữ liệu gồm các câu song ngữ Anh - Việt được lưu ở file en_sents và vi_sents.
- Dữ liệu được tải và chuẩn hóa (lowercase, tách từ).
- Xây dựng từ vựng (Vocab) cho cả hai ngôn ngữ bao gồm các token đặc biệt <pad>, <sos>, <eos>, <unk>.
- Câu được mã hóa thành chuỗi chỉ số tương ứng với từ vựng.
- Dữ liệu được tổ chức thành Dataset và DataLoader, với padding để đảm bảo batch có kích thước đồng nhất.

2.2. Mô hình RNN Seq2Seq.

- Sử dụng mô hình Seq2Seq với Encoder và Decoder đều là GRU nhiều lớp.
- Encoder nhận vào chuỗi đầu vào tiếng Anh, chuyển thành biểu diễn ẩn (hidden state).
- Decoder dựa vào trạng thái ẩn để sinh chuỗi đích tiếng Việt từng bước.
- Áp dụng kỹ thuật teacher forcing với tỷ lệ 0.5 trong quá trình huấn luyện.
- Loss function: CrossEntropyLoss với ignore index cho padding.
- Huấn luyện với 5 cấu hình siêu tham số khác nhau bao gồm embedding dimension, hidden dimension, số lớp GRU, batch size và learning rate.
- Theo dõi quá trình huấn luyện qua loss trung bình mỗi epoch.

2.3. Mô hình Transformer.

- Sử dụng kiến trúc Transformer theo chuẩn Vaswani et al. gồm các lớp Encoder-Decoder, multi-head self-attention và feed-forward.
- Các cấu hình thay đổi embedding size, số lượng lớp encoder và decoder, số head attention và kích thước lớp feed-forward.
- Mô hình được huấn luyện trong 20 epoch mỗi cấu hình.
- Theo dõi các chỉ số như train loss, validation loss, BLEU score bằng wandb.
- Kết quả BLEU trên validation đạt từ khoảng 0.31 đến 0.38 tùy cấu hình.
- Mô hình được lưu lại và có thể tải để sử dụng cho dịch câu mới.

2.4. Theo dõi và trực quan hóa.

Sử dụng Weights & Biases (wandb) để log tự động các thông số huấn luyện: loss, BLEU, thời gian, siêu tham số.

Dữ liệu wandb cho phép so sánh hiệu quả giữa các cấu hình nhanh chóng, trực quan.

Từ dữ liệu wandb, xác định các cấu hình có hiệu suất tốt nhất.

2.5. Giao diện demo.

- Xây dựng giao diện người dùng đơn giản bằng Gradio.
- Người dùng nhập câu tiếng Anh, hệ thống trả về câu dịch tiếng Việt do mô hình Transformer thực hiện.
- Sử dụng kỹ thuật greedy decoding để sinh câu dịch từ mô hình.
- Giao diện thân thiện, dễ dùng và có thể chạy trên máy cục bộ hoặc deploy trên web.

3. Kết quả.

3.1. Mô hình RNN Seq2Seq.

- Mô hình đơn giản, nhanh trong huấn luyện nhưng chất lượng dịch thấp hơn nhiều so với Transformer.
- Loss giảm đều trong quá trình huấn luyện nhưng không đạt được kết quả cao về độ chính xác hay BLEU.
- Thích hợp để làm nền tảng, so sánh.

3.2. Mô hình Transformer.

- Các cấu hình thử nghiệm với các embedding size từ 256 đến 512, số head từ 4 đến 8, số lớp encoder/decoder từ 2 đến 4.
- Kết quả BLEU tốt nhất đạt khoảng 0.38, thể hiện khả năng dịch khá tốt.

- Ví dụ cấu hình tốt:
- Embedding 512, 4 lớp encoder, 8 head attention, FFN 1024, đạt BLEU ~0.37-0.38.
- Loss huấn luyện giảm ổn định, validation loss thấp hơn so với mô hình RNN.
- Thời gian huấn luyện hợp lý (1-3 giờ cho mỗi cấu hình trên GPU).

3.3. Trực quan hóa và theo dõi.

Biểu đồ loss và BLEU rõ ràng qua wandb giúp dễ dàng so sánh và chọn lựa cấu hình tối ưu.

Tích hợp wandb hỗ trợ quản lý nhiều experiment song song.

3.4. Demo Gradio.

- Giao diện dịch tiếng Anh sang tiếng Việt hoạt động mượt mà, phản hồi nhanh.
- Người dùng có thể nhập bất kỳ câu tiếng Anh nào, nhận kết quả dịch với chất lượng cao nhờ mô hình Transformer.
- Demo hỗ trợ mở rộng, dễ dàng tích hợp vào các ứng dụng thực tế.

4. Đánh giá và nhận xét.

Ưu điểm:

- Transformer thể hiện hiệu quả vượt trội so với RNN truyền thống.
- Thử nghiệm nhiều cấu hình siêu tham số giúp tìm ra mô hình tối ưu.
- Theo dõi quá trình huấn luyện qua wandb hỗ trợ phân tích chuyên sâu.
- Giao diện demo tăng tính ứng dụng, tiện lợi cho người dùng cuối.

Hạn chế:

- Mô hình RNN chưa được tối ưu tốt cho các câu dài hoặc phức tạp.
- Các chỉ số BLEU vẫn còn dư địa cải thiện, có thể nâng cao bằng kỹ thuật tăng cường dữ liệu hoặc mô hình lớn hơn.
- Cần đánh giá thêm trên tập kiểm tra độc lập, tính toán độ lệch chuẩn để đảm bảo tính ổn định.

5. Hướng phát triển tiếp theo.

- Thử nghiệm thêm các mô hình SOTA khác như BERT-based Seq2Seq, mT5, hoặc các mô hình Transformer lớn hơn.
- Cải thiện tiền xử lý dữ liệu, thêm kỹ thuật tokenization nâng cao (byte pair encoding, sentencepiece).
- Tích hợp cơ chế beam search hoặc sampling để nâng cao chất lượng dịch.
- Triển khai demo trên môi trường web thực tế hoặc mobile app.
- Thu thập thêm tập dữ liệu lớn và đa dạng để nâng cao khả năng tổng quát hóa của mô hình.