Sentiment Analysis

*E-mail: pima.vn@gmail.com

Mô tả dự án

Dữ liệu văn bản là một trong những hình thức phổ biến nhất của dữ liệu, được sử dụng trong nhiều bài toán máy học phổ biến:

(1) *Phân loại văn bản (text classification)*: Dự đoán xem người sử dụng có hài lòng với sản phẩm hay không dựa vào reviews; phát hiện email spam.
(2) *Dịch tự động (machine translation)*
(3) *Trả lời câu hỏi tự động (question answering) và chatbot*
(4) *Tóm tắt văn bản (text summarization)*

Một trong những tính chất đặc thù của dữ liệu dạng văn bản đó chính là tính chất phụ thuộc của mỗi từ vào các từ trước đó. Giả sử chúng ta có câu "Tôi thích đọc", từ tiếp theo khả năng sẽ là "sách" hoặc "truyện", mà không phải là "thịt bò".
Tính chất này đưa đến cho chúng ta ý tưởng xây dựng một mô hình cho dữ liệu văn bản:

$$h_t = f(x_t, h_{t-1})$$

với h_t được gọi là vector "trạng thái ẩn" (hidden state) tại vị trí t, được tính phụ thuộc vào x_t là từ ở vị trí t và trạng thái ẩn h_{t-1} của từ trước $t-1$. Vector trạng thái ẩn h_{t-1} này có vai trò lưu lại thông tin của các từ trước nó, giúp chúng ta có thể mô hình hoá mối quan hệ giữa mỗi từ và các từ trước trong văn bản. Chúng ta có thể dùng những trạng thái ẩn này làm feature cho các bài toán máy học đã nêu ở trên.
Để biểu diễn hàm f, chúng ta có thể sử dụng một mạng nơ ron một lớp:

$$h_t = f(x_t, h_{t-1}) = W_x x_t + W_h h_{t-1} + b$$

Chú ý rằng các vector trọng số W_x, W_h, b được sử dụng chung để tính trạng thái ẩn cho mọi từ trong câu. Điều này đảm bảo rằng chúng ta có thể xử lý một câu có độ dài bất kì một cách đơn giản. Ta có thể gọi mô hình này là một **mạng nơ ron hồi quy (recurrent neural network)** truyền thống.
Tuy nhiên, trong thực tế, mạng nơ ron này thường gặp vấn đề trong việc mô hình những câu có độ dài tương đối lớn. Vấn đề này được giải quyết phần nào bởi mạng nơ ron Long Short-Term Memory của Hochreiter và Schmidhuber [4]. Ở dự án này, các bạn sẽ tìm hiểu về LSTM và ứng dụng nó vào bài toán **Sentiment Analysis** (nhận diện cảm xúc văn bản).

Câu hỏi gợi ý

(1) Trình bày một cách đơn giản để hiểu cách tối ưu mạng nơ ron hồi quy (gợi ý: so sánh với một mạng nơ ron có nhiều lớp!)

(2) Tìm hiểu các vấn đề gradient vanishing, gradient exploding của mạng nơron hồi quy truyền thống, và nghiên cứu tại sao những vấn đề này khiến mạng nơ ron hồi quy truyền thống gặp khó khăn trong việc mô hình các câu dài.

(3) Tìm hiểu định nghĩa của mạng LSTM, và giải thích (bằng toán hoặc bằng trực quan) cách các cơ chế cổng (gating mechanism) giúp giải quyết các vấn đề của mạng nơ ron hồi quy truyền thống.

(4) Ứng dụng RNN và LSTM vào bài toán phân loại sắc thái (positive/negative/neutral) của review (ví dụ: review trên foody). So sánh độ chính xác (accurracy), thời gian train (training time) của RNN và LSTM.

Một số từ khoá: Backpropagation Through Time (BPTT), Gradient Vanishing, Gradient Exploding, Long Short-Term Memory (LSTM)

Tham Khảo

[1] Các bài giảng PiMA 2019.

[2] Christopher Olah. Understanding LSTM Networks. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

[3] Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press, 2016. https://www.deeplearningbook.org/

[4] Sepp Hochreiter and Jurgen Schmidhuber.Long Short-Term Memory. https://www.bioinf.jku.at/publications/older/2604.pdf

[5] https://en.wikipedia.org/wiki/Recurrent_neural_network