Báo cáo VietAI – Final Assignment

Neural Machine Translation

Nguyễn Trung Thành

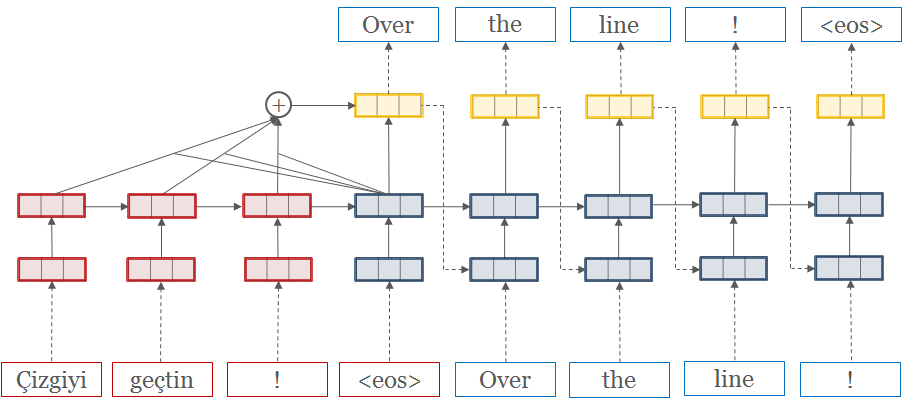
Tháng 10 năm 2019

1. Tổng quan về cấu trúc mô hình NMT và phương pháp đánh giá mô hình Machine Translation (Bleu Score)
   1. Tổng quan về cấu trúc mô hình NMT

Neural Machine Translation (NMT) là phương pháp học tập end-to-end để dịch tự động, với khả năng khắc phục nhiều điểm yếu của các phương pháp dịch máy truyền thống.

Mỗi ngôn ngữ đều có những đặc trưng riêng về ngữ pháp, cú pháp và ngữ nghĩa vì vậy phương pháp dịch máy truyền thống dựa trên việc dịch từng cụm từ (phrase-by-phrase) còn nhiều hạn chế, làm mất đi sự lưu loát của câu văn. Với NMT là một cách tiếp cận [dịch máy](https://vi.wikipedia.org/wiki/D%E1%BB%8Bch_t%E1%BB%B1_%C4%91%E1%BB%99ng" \o "Dịch tự động) (Machine Translation) sử dụng [mạng nơ-ron nhân tạo](https://vi.wikipedia.org/wiki/M%E1%BA%A1ng_n%C6%A1-ron_nh%C3%A2n_t%E1%BA%A1o" \o "Mạng nơ-ron nhân tạo) lớn để dự đoán chuỗi từ được dịch, bằng cách mô hình hóa toàn bộ các câu văn trong một mạng nơ-ron nhân tạo duy nhất. Với cách làm này, NMT đã đưa ra những kết quả tốt ngang ngửa với con người.

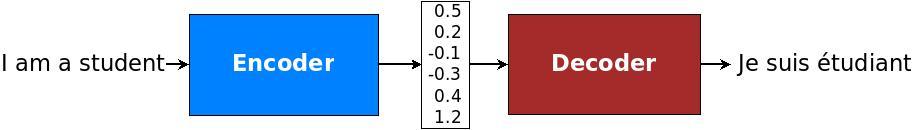
Kiến trúc của mô hình NMT gồm 2 thành phần chính là: mô hình Sequence-to-Sequence và cơ chế Attention.



1.1.1 Mô hình Sequence-to-Sequence

Sequence to Sequence Model (Seq2seq) là một mô hình Deep Learning với mục đích tạo ra một output sequence (câu đầu ra) từ một input sequence (câu đầu vào) mà độ dài của 2 sequences này có thể khác nhau. Seq2seq được giới thiệu bởi nhóm nghiên cứu của Google vào năm 2014 trong bài báo [Sequence to Sequence with Neural Networks](https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks).

Seq2seq gồm 2 phần chính là Encoder (bộ mã hoá) và Decoder (bộ giải mã). Cả hai thành phần này đều được hình thành từ các mạng Neural Networks, trong đó Encoder có nhiệm vụ chuyển đổi input sequence thành một representation với lower dimension còn Decoder có nhiệm vụ tạo ra output sequence từ representation của input sequence được tạo ra ở phần Encoder.

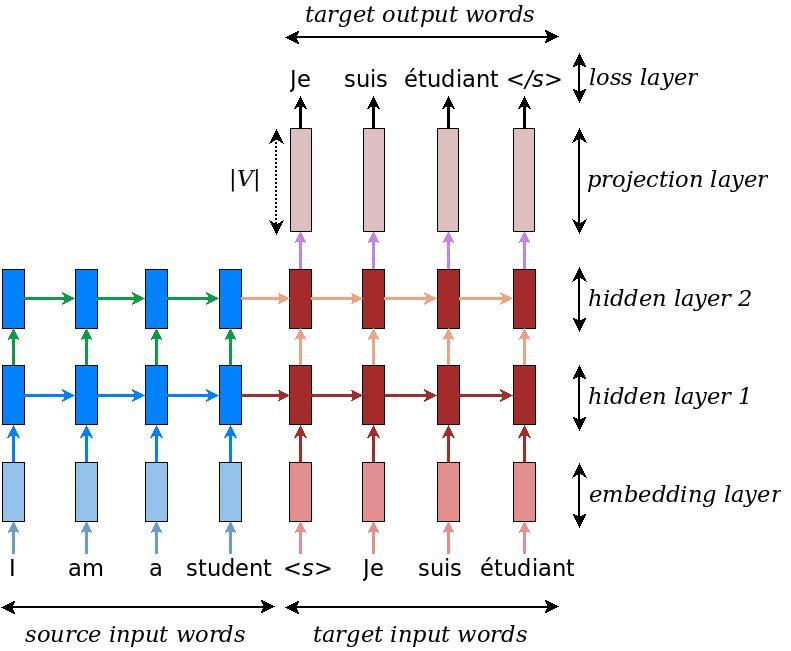


Hình 1: Kiến trúc Encoder-Decoder

Cụ thể, đầu tiên NMT sẽ đưa input sequence vào Encoder để nén ý nghĩa của câu tạo thành  ["thought" vector](https://www.theguardian.com/science/2015/may/21/google-a-step-closer-to-developing-machines-with-human-like-intelligence), một chuỗi các số đại diện sẽ chứa ý nghĩa của câu. Sau đó, Decoder xử lý để chuyển vector thành câu được dịch sang ngôn ngữ khác như minh họa trong hình 1.

Tùy từng bài toán cụ thể mà Encoder và Decoder sử dụng các kỹ thuật Deep Learning khác nhau. Ví dụ như trong Machine Translation thì Encoder thường là LSTM, GRU hoặc Bi-directional RNN, còn trong Image Captioning thì Encoder lại là CNN.

Trong báo cáo này, chúng ta sử dụng Seq2seq cho bài toán Machine Learning. Từ dữ liệu đầu vào là một sequence dưới dạng text, chúng ta sử dụng Embedding Layer để chuyển các từ này sang dạng Word Embedding rồi sử dụng RNN (thường là Bi-directional RNN) để tạo ra một representation của input sequence (trong hình bên dưới là <s>).



*Hình 2: Minh hoạ cấu trúc NMT*

Decoder cũng được tạo thành từ RNN và sử dụng output của Encoder làm dữ liệu đầu vào để tạo ra một output sequence. Tuy nhiên khác với Language Modeling, trong Machine Translation chúng ta phải chọn câu văn phù hợp nhất thay vì để RNN cell tạo ra từng từ một. Thông thường việc lựa chọn output sequence được thực hiện bởi các Search Algorithms với hai phương pháp chính sau:

– Greedy search : chọn từ có xác xuất cao nhất làm output của từng Cell. Ưu điểm của phương pháp này là có tốc độ nhanh, nhưng thường sẽ không tạo ra câu văn hợp lý nhất.

– Beam search: tại mỗi Decoding step, chúng ta chọn n-words (beam width) với xác xuất cao nhất. Ví dụ khi chúng ta chọn beam\_width=3 thì tại mỗi Decoding step, ta sẽ giữ lại 3 từ có xác xuất cao nhất rồi lấy từng từ một làm đầu vào cho Decoding step tiếp theo. Cứ như thế lặp lại cho đến khi ta gặp <EOS> đánh dấu việc kết thúc câu (xem thêm [tại đây](https://www.coursera.org/lecture/nlp-sequence-models/beam-search-4EtHZ)).

*(Chúng ta có thể coi Greedy search là một trường hợp của Beam search với việc sử dụng beam\_width=1).*

1.1.2 Vấn đề của mô hình Sequence-to-Sequence

Với Machine Translation thì cả Encoder và Decoder đều được tạo thành từ RNN cell (LSTM hoặc GRU). Về mặt lý thuyết thì LSTM và GRU có thể lưu trữ thông tin của một sequence có độ dài lớn. Tuy nhiên, trong thực tế việc sử dụng một vector representation thường không thể lưu trữ được toàn bộ thông tin của input sequence. Do đó, trong một số bài báo khoa học có trình bày một vài phương pháp giúp tăng độ chính xác cho hệ thống này như:

– Sử dụng Multi-layer với Bi-directional RNN

– Đảo ngược thứ tự của input sequence. Ví dụ như câu: ‘I enjoy eating’ thì thứ tự timestep được chuyển thành ‘eating enjoy I’

– Sử dụng input nhiều lần (feeding twice) nhưng vẫn giữ nguyên output (xem thêm tại đây)

Tuy nhiên, phương pháp được sử dụng nhiều nhất và làm tăng đáng kể độ chính xác của các hệ thống là sử dụng Attention Mechanism.

1.1.3 Cơ chế Attention

***Tài liệu tham khảo***

Neural Machine Translation (seq2seq) Tutorial - <https://github.com/tensorflow/nmt>

<https://github.com/huanpc/nmt_learning/tree/master/report>

<https://www.academia.edu/38225926/Neural_Machine_Translation_Overview>

<https://viblo.asia/p/machine-learning-attention-attention-attention-eW65GPJYKDO>

<https://viblo.asia/p/neural-machine-translation-mo-hinh-dich-may-va-danh-gia-mo-hinh-dich-may-Qbq5QkVRZD8>

<http://itechseeker.com/tutorials/nlp-with-deep-learning/ly-thuyet-chung/sequence-to-sequence-model/>

<http://itechseeker.com/tutorials/nlp-with-deep-learning/ly-thuyet-chung/attention-mechanism/>