Bài Tập - Neural Machine Translation

Ngày 19 tháng 10 năm 2019

Tóm tắt nội dung

Machine Translation (Dịch máy) là một bài toán trong Xử Lý Ngôn Ngữ Tự Nhiên (NLP) với mục đích dịch một câu/đoạn văn bản từ ngôn ngữ nguồn (source language) sang ngôn ngữ đích (target language). Việc áp dụng các mô hình trong Deep Learning như BiLSTM và Sequence-to-Sequence vào bài toán dịch máy đã đạt đạt được những kết quả đáng chú ý khi so sánh với các phương pháp dịch máy truyền thống. Trong bài tập này, các bạn sẽ sử dụng những kiến thức đã học về các mô hình Seuqnece-to-Sequence cho bài toán dịch máy.

1 Giới thiệu

Machine Translation nhằm giải quyết vấn đề dịch tự động các văn bản từ ngôn ngữ nguồn (source language) sang ngôn ngữ đích (target language). 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.

Trong thực tế, để dịch một đoạn văn bản, chúng ta sẽ đọc hiểu ý nghĩa trước khi tiến hành dịch thay vì tách đoạn thành từng cụm từ đơn lẻ. Neural Machine Translation là một phương pháp mô phỏng lại cách thức dịch này. Sử dụng mô hình Sequence-to-Sequence với một bộ mã hóa câu từ ngôn ngữ nguồn về dạng vector sau đó giải mã thành câu văn ở ngôn ngữ đích.

Trong bài tập này, học viên sẽ tìm hiểu về mô hình Neral Machine Translation của tác giả Lương Minh Thắng. Thông tin chi tiết về mô hình, mã nguồn cũng như cách thức huấn luyện mô hình được đăng tải trên trang github: **Neural Machine Translation (seq2seq) Tutorial**[1].

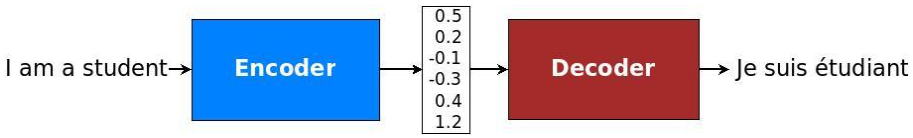
[1] https://github.com/tensorflow/nmt

2 Mục tiêu và các bước thực hiện

Thông qua việc tìm hiểu và huấn luyện lại mô hình Neural Machine Translation được giới thiệu ở trên, các học viên cần đạt được những mục tiêu sau:

2.1 Hiểu được kiến trúc của mô hình NMT

Mô hình Neural Machine Translation được hiện thực dựa trên kiến trúc Encoder - Decoder. Câu gốc được mã hóa bằng bộ Encoder để tạo ra vector đặc trưng, còn gọi là "Though"Vector. Sau đó bộ Decoder sẽ giải mã vector này về kết quả được dịch.



Hình 1: Kiến trúc tổng quan Encoder - Decoder. Nguồn: `https://github.com/tensorflow/nmt`

Thông tin về các lớp của mạng neural, hàm mất mác, cũng như các kỹ thuật nâng cao (attention, beam search, etc.) được mô tả cụ thể trong bài thực hành Neural Machine Translation (seq2seq) Tutorial. Học viên cần nắm được kiến trúc tổng quan của mô hình NMT cũng như cách thức hiện thực mô hình trên nền tảng Tensorflow thông qua mã nguồn của mô hình.

2.2 Huấn luyện mô hình NMT trên bộ dữ liệu Anh - Việt

IWSLT English-Vietnamese[2] là bộ dữ liệu bao gồm 133 nghìn cặp câu Anh - Việt từ TED talks, được cung cấp bởi *IWSLT Evaluation Campaign*. Đây là bộ dữ liệu tương đối nhỏ, thích hợp để học viên thử nghiệm việc huấn luyện mô hình, không đòi hỏi thời gian huấn luyện quá lớn và các yêu cầu cao về phần cứng.

Dữ liệu đầu vào cho mô hình NMT được xử lý thông qua **Data Input Pipeline**[3]. Học viên cần nắm được các bước tiền xử lý đối với dữ liệu dạng văn bản và cách thức đưa dữ liệu vào huấn luyện mô hình. Đồng thời có khả năng huấn luyện lại một mô hình dịch máy cho bộ ngôn ngữ Anh - Việt đạt kết quả tương tự kết quả được trình bày trong bài thực hành.

Để huấn luyện mô hình, học viên có thể sử dụng môi trường *Google Colab* hỗ trợ huấn luyện trên GPU. Chi tiết về cách sử dụng Google Colab được trình bày trong bài *Assigment 2 - RNN*.

[2] https://github.com/tensorflow/nmt#iwslt-english-vietnamese
[3] https://github.com/tensorflow/nmt#data-input-pipeline

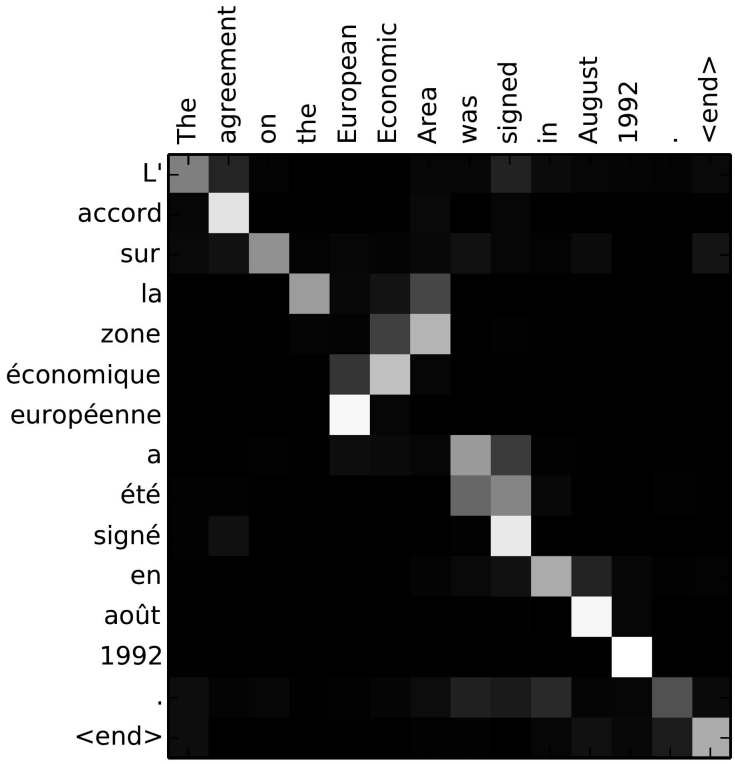
2.3 Tìm hiểu ảnh hưởng của các siêu tham số đối với kết quả huấn luyện

Trong các bài toán Deep Learning, việc lựa chọn siêu tham số sẽ có ảnh hưởng nhất định đến chất lượng của mô hình. Một số siêu tham số thường được quan tâm khi huấn luyện như tốc độ huấn luyện và giải thuật Gradient Descent.

Trong bài thực hành này, học viên cần tiến hành thí nghiệm với các bộ siêu tham số khác nhau, thông qua đó rút ra nhận xét về ảnh hưởng của các siêu tham số đối với kết quả huấn luyện.

2.4 Đánh giá chất lượng của kỹ thuật Attention

Ý tưởng của kỹ thuật Attention trong mô hình NMT là thể hiện mối liên hệ giữa các từ trong câu văn nguồn với từ tương ứng trong câu văn đích. Việc sử dụng kỹ thuật Attention sẽ giúp tăng chất lượng dịch đối với các câu văn dài.



Hình 2: Ví dụ về biểu diễn ma trận Attention. Nguồn: Bahdanau et al., 2015

Trong bài tập thực hành này, học viên cần nằm được nguyên lý hoạt động của kỹ thuật Attention, ảnh hưởng của Attention đối với chất lượng dịch, cũng như biểu diễn được ma trận Attetion và đánh giá chất lượng của ma trận Attention trên các cặp câu ví dụ.

3 Yêu cầu và đánh giá

Học viên cần trình bày các nội dung tìm hiểu về mô hình NMT cũng như kết quả huấn luyện mô hình trên tập dữ liệu IWSLT English-Vietnamese dưới dạng một bài báo cáo. Cấu trúc bài báo cáo bao gồm:

- Mô tả tổng quan về cấu trúc mô hình NMT và phương pháp đánh giá mô hình Machine Translation (Bleu Score)
- Trình bày kết quả và nhận xét về ảnh hưởng của các siêu tham số đối với mô hình NMT. Các thông số cần báo cáo bao gồm thời gian huấn luyện và kết quả của mô hình trên các bộ siêu tham số khác nhau: SGD với learning rate 1.0 và Adams với learning rate 0.001.
- Biểu diễn ma trận Attention cho các cặp câu Anh - Việt và nhận xét về chất lượng của ma trận Attention.

—— Chúc các bạn hoàn thành tốt ——