Report of Neural Machine Translation Model

Huan Phan

VietAI Course 3's student, HCM, Vietnam

Abstract

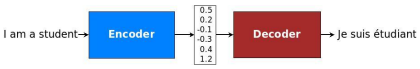
XYZ

Keywords: Neural Machine Translation, Seq2Seq, Attention

1. Overview of Neural Machine Translation

1.1. Encoder - Decoder Architect

Trong mô hình seq2seq dùng cho bài toán NMT (Neural Machine Translation) bao gồm 2 mạng RNN chính: Encoder và Decoder. Encoder với đầu vào là câu ở ngôn ngữ gốc, đầu ra tại layer cuối cùng của Encoder gọi là 1 context vector. Với ý nghĩa lượng thông tin từ câu của Encoder sẽ được tóm gọn lại trong 1 vector đầu ra cuối cùng. Từ đó, Decoder dùng chính context vector đó, cùng với hidden state và từ trước đó để predict từ tiếp theo tại decoder qua từng timestep.

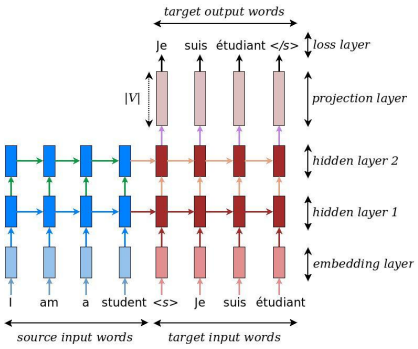


Hình 1: Encoder-decoder architecture – example of a general approach for NMT. An encoder converts a source sentence into a "meaning" vector which is passed through a decoder to produce a translation.

1.2. Neural Machine Translation Architect

1.3. Attention Mechanise

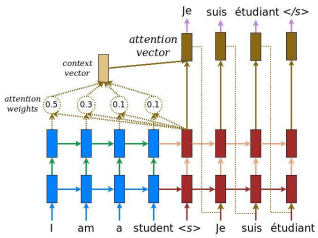
We conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder–decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly[1]



Hình 2: Neural machine translation – example of a deep recurrent architecture proposed by for translating a source sentence "I am a student" into a target sentence "Je suis étudiant". Here, "<s>" marks the start of the decoding process while "</s>" tells the decoder to stop.

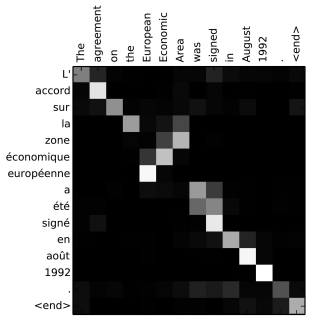
Việc encode toàn bộ thông tin từ source vào 1 vector cố định khiến việc mô hình khi thực hiện trên các câu dài (long sentence) không thực sự tốt, mặc dù sử dụng LSTM (BiLSTM, GRU) để khắc phục điểm yếu của mạng RNN truyền thống với hiện tượng Vanishing Gradient, nhưng như thế có vẻ vẫn chưa đủ, đặc biệt đối với những câu dài hơn những câu trong training data. Từ đó, trong paper, tác giả Bahdanau đề xuất 1 cơ chế cho phép mô hình có thể chú trọng vào những phần quan trọng (word liên kết với word từ source đến target), và thay vì chỉ sử dụng context layer được tạo ra từ layer cuối cùng của Encoder, tác giả sử dụng tất cả các output của từng cell qua từng timestep, kết hợp với hidden state của từng cell để "tổng hợp" ra 1 context vector (attention vector) và dùng nó làm đầu vào cho từng cell trong Decoder.

Cơ chế "tổng hợp" Attention trong paper của tác giả Bahdanau: Align and Jointly model (Additive Attention).



Hình 3: Attention mechanism – example of an attention-based NMT system as described in (Luong et al., 2015) . We highlight in detail the first step of the attention computation. For clarity, we don't show the embedding and projection layers in Figure (2).

Ma trận bất đối xứng (confusion matrix) được tạo ra bởi alignment score, thể hiện mức độ tương quan correlation giữa source và target.



Hình 4: Attention visualization – example of the alignments between source and target sentences. Image is taken from (Bahdanau et al., 2015).

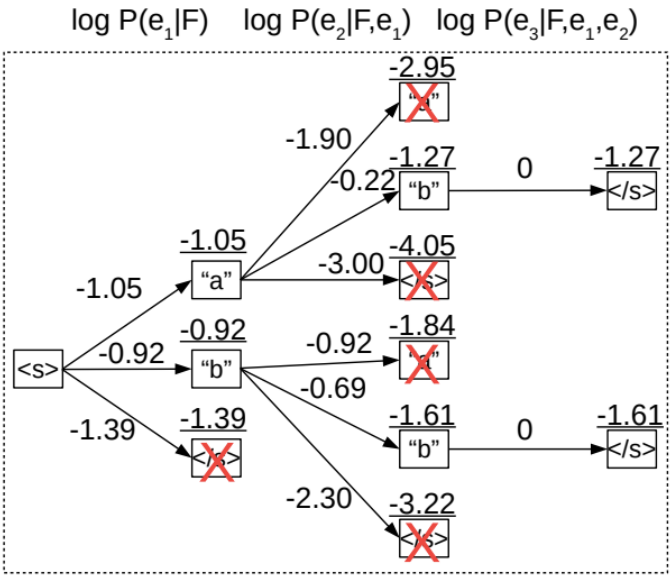
1.4. (SGD, Adam)

1.5. Inference mode

Được sử dụng trong quá trình decoding tại module decoder.

Greedy: Câu nguồn được module encoder encode thành encoder_state để khởi tạo giá trị đầu vào cho module decoder. Qúa trình decoding (dịch) bắt đầu khi decoder đọc được ký tự bắt đầu câu <*s*>.Tại mỗi thời điểm t, ouput của mạng RNN là một vector V chiều (V là kích thước bộ vocab), với mỗi phần tử của vector là các giá trị xác suất tương ứng với các từ trong vocab. Chiến thuật Greedy (tham lam) là chọn từ tương ứng với vị trí của xác suất cao nhất, từ này sẽ là output của NMT tại thời điểm t, và sẽ thành đầu vào cho decoder tại thời điểm tiếp theo.

Beam search: Beam search decoder có thể làm tăng hiệu quả của mô hình. Ý tưởng của beam search là không chọn duy nhất một từ có giá trị xác suất dự đoán cao nhất (như greedy), mà giữ lại một tập hợp nhỏ top các từ có giá trị xác suất cao nhất. Tập các ứng cử viên này trở thành đầu vào cho decoder, tạo ra không gian output lớn hơn cho bước tiếp theo. Kích thước của tập hợp này gọi là beam width, thông thường có giá trị bằng 10, là đủ hiệu quả. Phương pháp này cho phép bước sau chọn ra ứng cử viên có xác suất đồng thời cao nhất từ các ứng cử viên đã chọn được từ bước trước, giảm thiểu lỗi chọn sai vì top các ứng cử viên từ bước trước đã được giữ lại.[2]



Hình 5: Ví dụ beam search với k=2. Trên các mũi tên là log probability của từng từ $P(e_t|F, e_1^{t-1})$. Trên các node là log probability tổng của các hypothesis tính đến node đó.

2. Measure performance with BLEU score

$$P(e_t|F, e_1^{t-1}) \tag{1}$$

3. Training with different hyper-parameters

3.1. IWSLT English-Vietnamese dataset

Train: 133K examples, vocab=vocab.(vi|en), train=train.(vi|en) dev=tst2012.(vi|en), test=tst2013.(vi|en)

3.2. Data Input Pipeline

https://github.com/tensorflow/nmtdata-input-pipeline

3.3. Impact of choosing hyper-parameters

3.3.1. Training details

I train 2-layer LSTMs of 512 units with bidirectional encoder (i.e., 1 bidirectional layers for the encoder), embedding dim is 512. LuongAttention (scale=True) is used together with dropout *keep_prob* of 0.8. All parameters are uniformly. We use SGD with learning rate 1.0 as follows: train for 12K steps (12 epochs); after 8K steps, we start halving learning rate every 1K step.

Training system: Tesla K80

3.3.2. Result with VI - EN translation model

Parameter	**Time**	**BLEU score**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Bảng 1: VI - EN performances

3.3.3. Result with EN - VI translation model

Experiment	**Training Time**	**tst2012 (dev) BLEU**	**test2013 (test) BLEU**
Treatment 1	0.0003262	0.562	1
Treatment 2	0.0015681	0.910	1
Treatment 3	0.0009271	0.296	1

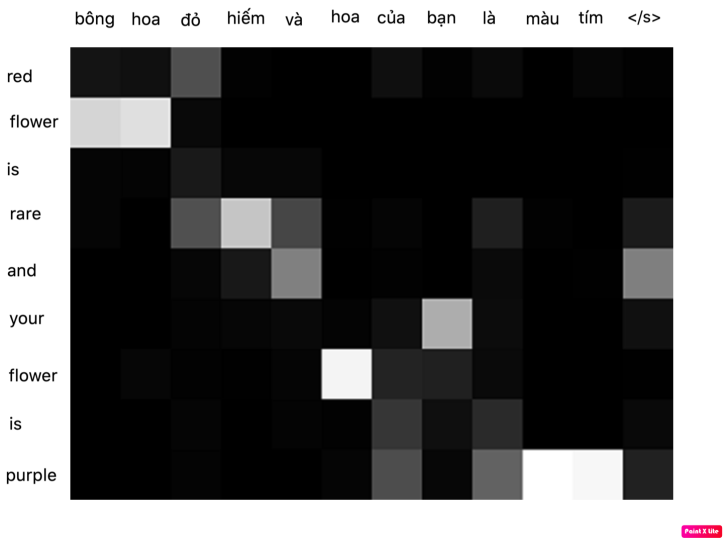
Bảng 2: VI - EN performances

3.3.4. Conclusion

4. Result visualization with Attention Matrix

4.1. Test 1

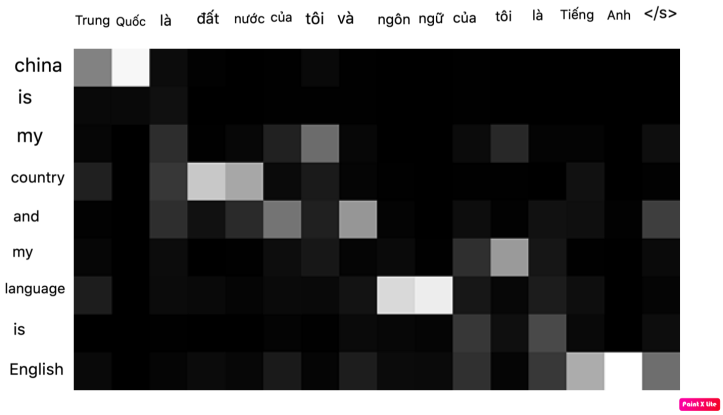
- red flower is rare and your flower is purple
- bông hoa đỏ hiếm và hoa của bạn là màu tím



Hình 6: Attention alignment giữa câu nguồn "red flower is rare and your flower is purple" và câu đích "bông hoa đỏ hiếm và hoa của bạn là màu tím".

4.2. Test 2

- China is my country and my language is English
- Trung Quốc là đất nước của tôi và ngôn ngữ của tôi là tiếng Anh



Hình 7: Attention alignment giữa câu nguồn "China is my country and my language is English" và câu đích "Trung Quốc là đất nước của tôi và ngôn ngữ của tôi là tiếng Anh".

Tài liệu

[1] J. M. Smith, A. B. Jones, Book Title, Publisher, 7th edition, 2012.

[2] Graham, Neural Machine Translation and sequence-to-sequence models: A tutorial (????).