Học Tuần Tự Theo Tuần Tự Với Mạng Nơ-Ron

Loc PV. Nguyen*, Khoi XM. Nguyen, Phuong H. Nguyen, Hoang N. Dang

Faculty of Information Technology, FPT School Of Business And Technology, Ho Chi Minh city, Vietnam

Tóm tắt

Các mô hình mạng học sâu (Deep Neural Networks - DNNs) là những mô hình mạnh mẽ đã đạt được hiệu suất xuất sắc trong các nhiệm vụ học tập phức tạp. Mặc dù DNNs hoạt động tốt khi nào có sẵn số lượng lớn các tập huấn luyện gán nhãn, nhưng chúng không thể được sử dụng để ánh xạ chuỗi thành chuỗi. Bài báo này trình bày một cách tiếp cận tổng quát từ đầu đến cuối để học trình tự tạo ra các giả định tối thiểu về cấu trúc trình tự. Trong bài báo, tác giả đã sử dụng phương pháp mạng nhiều lớp bộ nhớ ngắn hạn dài (Long Short-Term Memory - LSTM) để ánh xạ chuỗi đầu vào thành một vectơ có chiều dài cố định và sau đó là một LSTM sâu khác để giải mã chuỗi mục tiêu từ vectơ đó. Mục tiêu chính là dịch từ tiếng Anh sang tiếng Pháp từ tập dữ liệu WMT'14, các bản dịch do LSTM tạo ra đạt được điểm BLEU là 34,8 trên toàn bộ tập thử nghiệm, trong đó điểm số BLEU của LSTM bị phạt đối với các từ không nằm trong tập từ vựng. Ngoài ra, LSTM không gặp khó khăn với các câu dài. Để so sánh, hệ thống dịch dựa trên cụm từ (phrase-based) SMT đạt được điểm BLEU là 33,3 trên cùng một tập dữ liệu. Trong khi đó, sử dụng LSTM để đánh giá lại 1000 giả thuyết được tạo ra bởi hệ thống SMT nói trên, điểm BLEU của nó tăng lên 36,5, gần với kết quả tốt nhất trước đó cho nhiệm vụ này. LSTM cũng học các cách biểu diễn câu và cụm từ hợp lý nhạy cảm với trật tự từ và tương đối bất biến đối với giọng chủ động và giọng bị động. Cuối cùng, việc đảo ngược thứ tự của các từ trong tất cả các câu nguồn (nhưng không phải câu đích) đã cải thiện đáng kể hiệu suất của LSTM, bởi vì làm như vậy đã tạo ra nhiều phụ thuộc ngắn

*Corresponding author

Email addresses: loc20mse23026@fsb.edu.vn (Loc PV. Nguyen), khoi20mse23024@fsb.edu.vn (Khoi XM. Nguyen), phuong20mse23020@fsb.edu.vn (Phuong H. Nguyen), hoang20mse23030@fsb.edu.vn (Hoang N. Dang)

hạn giữa câu nguồn và câu đích khiến vấn đề tối ưu hóa trở nên dễ dàng hơn.

Từ khóa: Mô hình học sâu, Mạng bộ nhớ gần-xa, Trí tuệ nhân tạo, Học máy, Hệ thống dựa trên cụm từ, Mạng nơ-ron hồi quy.

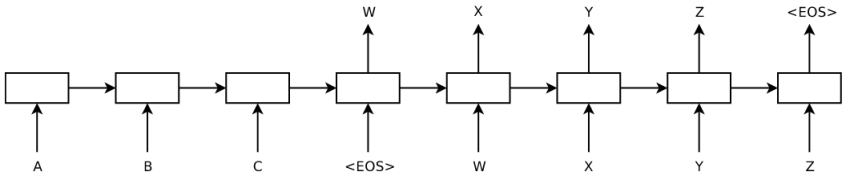
---

1. Giới thiệu

Các mạng học sâu (DNNs) là mô hình học máy cực kỳ mạnh mẽ đạt được hiệu suất tuyệt vời đối với các vấn đề khó khăn như nhận dạng giọng nói (Hinton et al., 2012; Dahl et al., 2012) và nhận dạng đối tượng trực quan (Krizhevsky et al., 2017; Ciresan et al., 2012; LeCun et al., 1998; Le et al., 2012). DNNs rất mạnh mẽ vì chúng có thể thực hiện tính toán song song tùy ý với một vài bước tối thiểu nhất. Một ví dụ đáng ngạc nhiên về sức mạnh của DNNs là khả năng sắp xếp N số bit N của chúng chỉ bằng cách sử dụng 2 lớp ẩn có kích thước bậc hai (Razborov, 1992). Vì vậy, trong khi mạng nơ-ron có liên quan đến các mô hình thống kê thông thường, chúng lại học phép tính một cách phức tạp. Hơn nữa, các mạng DNNs lớn được huấn luyện lan truyền ngược có giám sát bất cứ khi nào tập huấn luyện được gắn nhãn có đủ thông tin để chỉ định các tham số của mạng. Do đó, nếu tồn tại một tham số hiệu chỉnh của một mạng DNNs lớn mà mạng đó đạt được kết quả tốt (ví dụ: vì con người có thể giải quyết công việc rất nhanh), thì lan truyền ngược có giám sát sẽ tìm ra các tham số này và giải quyết vấn đề.

Mặc dù DNNs có tính linh hoạt, nhưng DNNs chỉ có thể được áp dụng cho các vấn đề mà đầu vào và mục tiêu đầu ra có thể được mã hóa hợp lý bằng các vectơ có chiều cố định. Đó là một hạn chế đáng kể, vì nhiều vấn đề quan trọng được thể hiện tốt nhất với các trình tự mà độ dài của nó không được biết trước. Ví dụ, nhận dạng giọng nói và dịch máy là các vấn đề tuần tự. Tương tự như vậy, trả lời câu hỏi cũng có thể được coi là ánh xạ một chuỗi các từ đại diện cho câu hỏi với một chuỗi các từ đại diện cho câu trả lời. Do đó, rõ ràng rằng một phương pháp độc lập với miền học cách ánh xạ các chuỗi thành các chuỗi sẽ hữu ích.

Các chuỗi tuần tự đặt ra một thách thức đối với các DNNs vì chúng yêu cầu rằng kích thước của các đầu vào và đầu ra phải được biết và cố định. Bài nghiên cứu này chỉ ra một ứng dụng đơn giản của kiến trúc Bộ nhớ Gần-Xa (LSTM) (Hochreiter and Schmidhuber, 1997) có thể giải quyết các vấn đề tuần tự tổng quát. Ý tưởng là sử dụng một LSTM để đọc chuỗi đầu vào, một bước thời gian tại một thời điểm, để biểu diễn vectơ có chiều cố định



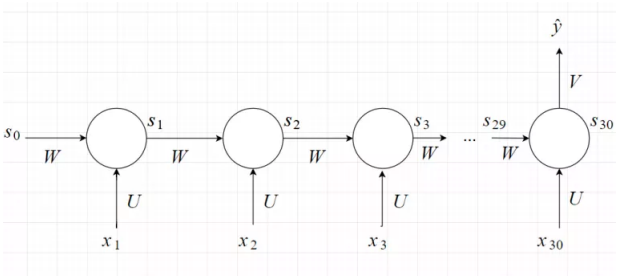
Hình 1: Mô hình đọc chuỗi đầu vào ABC và tạo chuỗi đầu ra WXYZ.

lớn và sau đó sử dụng một LSTM khác để trích xuất chuỗi đầu ra từ vectơ đó. LSTM thứ hai về cơ bản là một mô hình ngôn ngữ mạng nơ-ron hồi quy (Mikolov et al., 2010) ngoại trừ việc nó được điều chỉnh dựa trên chuỗi tuần tự đầu vào. Khả năng học thành công trên dữ liệu có phạm vi dài phụ thuộc vào thời gian của LSTM khiến nó trở thành lựa chọn đương nhiên cho ứng dụng này do độ trễ thời gian đáng kể giữa đầu vào và đầu ra tương ứng của chúng.

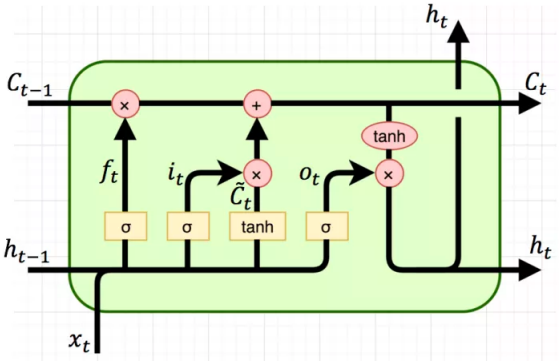
2. Hướng tiếp cận

Đã có một số báo cáo tương tự chỉ ra vấn đề chung của phương pháp học trình tự với mạng nơ-ron. Cách tiếp cận trong bài báo này thì gần với (Kalchbrenner and Blunsom, 2013), người đầu tiên nghiên cứu giải thuật ánh xạ toàn bộ câu đầu vào thành vectơ và tương tự như (Cho et al., 2014) mặc dù sau này chỉ được sử dụng để phục hồi các giả thuyết được tạo ra bởi một hệ thống dựa trên cụm từ. Graves (Graves, 2013) đã giới thiệu kỹ thuật attention cho phép mạng nơ-ron tập trung vào các phần khác nhau trong đầu vào của chúng và một biến thể của ý tưởng này đã được Bahdanau áp dụng thành công cho việc dịch máy (Bahdanau et al., 2019). Phân loại trình tự kết nối là một kỹ thuật phổ biến khác để ánh xạ các trình tự thành các trình tự có mạng nơron, nhưng nó giả định một sự liên kết đơn phương giữa các đầu vào và đầu ra (Graves et al., 2006).

Mô hình đề xuất dựa trên mô hình mạng nơ-ron sâu LSTM, đây là một dạng đặc biệt của RNN (Recurrent Neural Network - Mạng nơ-ron hồi quy). LSTM được giới thiệu bởi (Hochreiter and Schmidhuber, 1997) nhằm giải quyết các bài toán về phụ thuộc xa (long-term dependency). Ý tưởng là sử dụng một LSTM để đọc chuỗi đầu vào, một bước thời gian tại một thời điểm, để biểu diễn vectơ có chiều cố định lớn và sau đó sử dụng một LSTM khác để trích xuất chuỗi đầu ra từ vectơ đó. LSTM thứ hai về cơ bản là một mô hình ngôn ngữ mạng nơ-ron hồi quy (Mikolov et al., 2010) ngoại trừ việc nó



Hình 2: Mô hình RNN



Hình 3: Mô hình LSTM

được điều chỉnh dựa trên chuỗi tuần tự đầu vào. Khả năng học thành công trên dữ liệu có phạm vi dài phụ thuộc vào thời gian của LSTM khiến nó trở thành lựa chọn đương nhiên cho ứng dụng này do độ trễ thời gian đáng kể giữa đầu vào và đầu ra tương ứng của chúng.

3. Phương pháp

Mô hình mạng nơ-ron hồi quy (Recurrent Neural Network, viết tắt là RNN) (Werbos, 1990; Rumelhart et al., 1986) là sự tổng quát hóa tự nhiên của các mạng nơ-ron truyền thẳng tới các chuỗi. Trong khoảng 5-6 năm gần đây, RNN được ứng dụng rộng rãi trong ngành NLP và thu được những thành tựu lớn. Mạng RNN mô hình hóa được bản chất của dữ liệu trong NLP (có đặc tính chuỗi và các thành phần như từ, cụm từ trong dữ liệu phụ thuộc lẫn nhau). Có thể nói việc áp dụng mạng RNN là một bước đột phá trong ngành NLP.

Cho trước 1 chuỗi $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$, một mạng RNN tiêu chuẩn tính toán

thứ tự kết quả đầu ra $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_T\}$ bằng cách lặp đi lặp lại phương trình sau:

$$\mathbf{h}_t = sigm\left(W^{hx}\mathbf{x}_t + W^{hh}\mathbf{h}_{t-1}\right) \tag{1}$$

$$\mathbf{y}_t = W^{yh}\mathbf{h}_t \tag{2}$$

RNN có thể dễ dàng ánh xạ chuỗi thành chuỗi bất cứ khi nào sự liên kết giữa các thông lượng đầu vào được biết trước. Tuy nhiên, phương pháp này không thực sự hiệu quả để áp dụng cho một RNN khi giải quyết các vấn đề mà đầu vào và chuỗi đầu ra có độ dài khác nhau cùng với mối quan hệ vào - ra phức tạp và không đơn điệu.

Chiến lược đơn giản nhất để học tuần tự chung là ánh xạ tuần tự đầu vào tới một trình lưu trữ có kích thước cố định bằng cách sử dụng một RNN và sau đó ánh xạ vectơ tới chuỗi mục tiêu bằng một RNN khác (cách tiếp cận này cũng đã được thực hiện bởi (Cho et al., 2014)). Mặc dù về nguyên tắc, nó có thể hoạt động vì RNN được cung cấp tất cả các thông tin liên quan, nhưng sẽ rất khó để đào tạo các RNN do các sự phụ thuộc xa (Hình 1). Tuy nhiên, Bộ nhớ dài-ngắn (LSTM) (Hochreiter and Schmidhuber, 1997) được biết là có khả năng tìm hiểu các vấn đề về phụ thuộc thời gian trong phạm vi dài, vì vậy LSTM có thể thành công trong cài đặt này.

Mục tiêu của LSTM là ước tính xác suất có điều kiện $\mathrm{P}(\mathbf{y}_{1:T'} \mid \mathbf{x}_{1:T})$ với $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ là chuỗi đầu vào và $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T'}\}$ là chuỗi đầu ra tương ứng với chiều dài T' có thể khác với T. LSTM tính xác suất có điều kiện này bằng cách đầu tiên lấy biểu diễn chiều cố định v của chuỗi đầu vào $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T'}\}$ cho bởi trạng thái ẩn cuối cùng của LSTM, rồi sau đó tính toán xác suất của $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T'}\}$ với công thức tiêu chuẩn LSTM-LM với trạng thái ẩn ban đầu được đặc trưng bởi v của chuỗi đầu vào $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{T'}\}$:

$$P\left(\mathbf{y}_{1:T'} \mid \mathbf{x}_{1:T}\right) = \prod_{t=1}^{T'} P\left(\mathbf{y}_t \mid v, \mathbf{y}_{1:t-1}\right) \tag{3}$$

4. Ứng dụng

Ngày nay, các ứng dụng của mô hình tuần tự theo tuần tự đã trở nên rất phổ biến. Không còn chỉ gói gọn trong các ứng dụng liên quan đến Dịch Máy, mô hình tuần tự theo tuần tự còn được trong ứng dụng sau:

4.1. Trả lời tự động (Dialog). Một trong những ứng dụng thú vị của tuần tự theo tuần tự là Hội Thoại, hay Hệ Thống Trả Lời Tự Động. Đầu vào của Hệ Thống sẽ là yêu cầu hoặc từ phía người dùng F (ví dụ: hỏi "Bạn tên gì"), đầu ra của hệ thống sẽ là phản hồi của hệ thống (E), như trong hàm được giới thiệu dưới đây :

$$\hat{E} = \underset{\mathbf{E}}{argmax}\, logP(E|F) - \lambda logP(E)$$

Trong đó, số hạng thứ nhất

$$\underset{\mathbf{E}}{argmax}\, logP(E|F)$$

là một hàm giải mã (decoded) thông thường của một mô hình seq2seq trong khi số hạng thứ 2

$$\lambda logP(E)$$

được sử dụng để tăng tính đa đạng của đầu ra, bằng cách "phạt" những kết quả đầu ra mà ít tính tương quan với đầu vào E(ví dụ: trả lời: "Tôi không biết").

4.2. Tóm tắt nội dung tự động (Summarization). : Việc tóm tắt nội dung một cách tự động được thực hiện thông qua một số lớp như sau:

Rút gọn câu: Rút gọn và ít làm thay đổi ngữ nghĩa của một câu đơn.

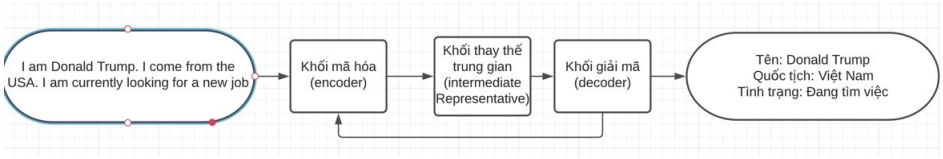
Rút gọn một đoạn văn bản:thực hiện thông qua rút gọn lần lượt các câu trong văn bản.

Rút gọn nhiều đoạn văn bản:thực hiện bằng cách rút gọn nội dung trong nhiều văn bản đơn lẻ thành một văn bản rút gọn duy nhất.

Kỹ thuật này còn được gọi là Tóm tắt nội dung theo phương thức trừu tượng (Abstractive Summarization)

4.3. Tạo diễn giải tự động (Paraphrase Generation). : Với mô hình diễn giải tự động, tương ứng với mỗi câu/ngữ cảnh ở đầu vào sẽ cho ra một câu/ngữ cảnh ở đầu ra có cùng ý nghĩa nhưng khác biệt về từ vựng và cú pháp.

4.4. Mô hình hóa các cấu trúc dữ liệu (Model of Structured Data). : Kỹ thuật seq2seq còn được sử dụng để biến dòng dữ liệu không có cấu trúc (unstructured data) ở đầu vào thành dữ liệu có cấu trúc ở đầu ra theo dạng cây, dạng bảng hay đồ thị. Ví dụ sau cho ta thấy điều đó:



Hình 4: Mô hình hóa các cấu trúc dữ liệu

5. Kết luận

Trong nghiên cứu này, bài báo đã chỉ ra rằng một mô hình LSTM sâu và lớn, có vốn từ vựng hạn chế và hầu như không có giả định nào về cấu trúc vấn đề vẫn có thể hoạt động tốt hơn một hệ thống SMT cổ điển có vốn từ vựng là không giới hạn đối với một nhiệm vụ MT quy mô lớn. Sự thành công của phương pháp đơn giản này cho thấy rằng nó sẽ giải quyết tốt nhiều vấn đề học tập theo trình tự khác, miễn là chúng có đủ dữ liệu đào tạo. Quan trọng nhất, bài báo đã chứng minh rằng một hướng tiếp cận đơn giản, dễ hiểu và tương đối không được tối ưu hóa có thể hoạt động tốt hơn hệ thống SMT, do đó, các công việc tiếp theo có khả năng dẫn đến độ chính xác dịch thuật lớn hơn. Những kết quả này cho thấy rằng cách tiếp cận của tác giả bài báo có thể sẽ giải quyết tốt các vấn đề chuỗi tuần tự đầy thách thức khác.

Tài liệu

Bahdanau, D., Cho, K., and Bengio, Y. (2019). Neural machine translation by jointly learning to align and translate.

Cho, K., van Merrienboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*.

Ciresan, D. C., Meier, U., and Schmidhuber, J. (2012). Multi-column deep neural networks for image classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3642–3649. IEEE Computer Society.

Dahl, G. E., Yu, D., Deng, L., and Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Trans. Speech Audio Process.*, 20(1):30–42.

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1700–1709.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90.

Le, Q. V., Ranzato, M., Monga, R., Devin, M., Corrado, G., Chen, K., Dean, J., and Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Kobayashi, T., Hirose, K., and Nakamura, S., editors, *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048. ISCA.

Razborov, A. A. (1992). On small depth threshold circuits. In Nurmi, O. and Ukkonen, E., editors, *Algorithm Theory - SWAT '92, Third Scandinavian Workshop on Algorithm Theory, Helsinki, Finland, July 8-10, 1992, Proceedings*, volume 621 of *Lecture Notes in Computer Science*, pages 42–52. Springer.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560.