Các mô hình tiền huấn luyện trong xử lý ngôn ngữ tự nhiên

Loc PV. Nguyen
FPT University Global Education
Ho Chi Minh City, Vietnam
loc20mse23026@fsb.edu.vn

Khoi XM. Nguyen
FPT University Global Education
Ho Chi Minh City, Vietnam
khoi20mse23024@fsb.edu.vn

Phuong H. Nguyen
FPT University Global Education
Ho Chi Minh City, Vietnam
phuong20mse23020@fsb.edu.vn

Hoang N. Dang
FPT University Global Education
Ho Chi Minh City, Vietnam
hoang20mse23030@fsb.edu.vn

***Tóm tắt nội dung*—Trong bài khảo sát này chúng tôi sẽ cung cấp cái nhìn toàn diện về thuật ngữ Tiền Mô Hình Đào Tạo trong Xử Lý Ngôn Ngữ Tự Nhiên (NLP). Mục tiêu của bài khảo sát này là tìm hiểu, nghiên cứu và so sánh tính hiệu quả của kỹ thuật này so với những phương pháp trước kia dựa trên những quy tắc chung.**

***Index Terms*—Xử lý ngôn ngữ tự nhiên, Mô hình tiền huấn luyện, Trí tuệ nhân tạo, Học máy.**

I. Giới thiệu

Phương pháp Học Chuyển Tiếp (Transfer Learning) là một phương pháp phổ biến trong thị giác máy tính cũng như xử lý ngôn ngữ tự nhiên và nhiều ứng dụng học máy khác. Học chuyển tiếp là một cách tiếp cận trong học sâu (và học máy), nơi kiến thức được chuyển giao từ mô hình này sang mô hình khác.

Với phương pháp học chuyển tiếp, thay vì bắt đầu quá trình huấn luyện (Training) từ đầu, ta có thể bắt đầu học từ các mô hình tiền huấn luyện (Pre-trained model) đã đạt được khi giải quyết một vấn đề khác. Bằng cách này, ta có thể tận dụng những kiến thức (features) đã học trước đó và tránh bắt đầu lại từ đầu.

Mô hình tiền huấn luyện (Pre-trained model) là một mô hình đã được đào tạo trên một tập dữ liệu chuẩn và đủ lớn để giải quyết một vấn đề tương tự như vấn đề mà chúng ta muốn giải quyết (như xử lý ngôn ngữ tự nhiên..). Do chi phí để huấn luyện các model rất tốn kém, nên thông thường người ta sẽ sử dụng các model từ các nguồn đã được public trước đó (ví dụ: BERT, PhoBERT, Underthesea, VGG, Inception, MobileNet,. . .).

Với sự phát triển của học sâu, các mạng nơ-ron khác nhau đã được sử dụng rộng rãi để giải quyết các bài toán NLP , chẳng hạn như mạng nơ-ron tích chập (Convolutional Neural Network) [1]–[3], mạng nơ-ron hồi quy (Recurrent Neural Network) [4], [5], mạng nơ-ron đồ thị (Graph Neural Network) [6]–[8]. Các phương pháp NLP không sử dụng mạng nơ-ron thần kinh thường chủ yếu dựa vào các tính năng được tạo thủ công rời rạc, trong khi các phương pháp thần kinh thường sử dụng các vectơ có chiều thấp và dày đặc (hay còn gọi là biểu diễn phân phối) để thể hiện ngầm định các đặc điểm ngữ nghĩa cú pháp của ngôn ngữ. Những đại diện này được học trong các nhiệm vụ NLP cụ thể. Do đó, các phương pháp thần kinh giúp mọi người dễ dàng phát triển các hệ thống NLP khác nhau.

Mặc dù mô hình mạng nơ-ron thần kinh cho các tác vụ NLP khá thành công, nhưng việc cải thiện hiệu suất có thể ít hơn so với lĩnh vực Thị giác máy tính (Computer Vision). Lý do chính là tập dữ liệu hiện tại cho hầu hết các tác vụ NLP được giám sát khá nhỏ (ngoại trừ dịch máy). Các mạng thần kinh sâu thường có một số lượng lớn các tham số, điều này làm cho chúng trang bị quá mức cho các dữ liệu huấn luyện nhỏ này và không tổng quát hóa tốt trong thực tế. Do đó, các mô hình thần kinh ban đầu cho nhiều tác vụ NLP tương đối nông và thường chỉ bao gồm 1/3 lớp thần kinh.

Gần đây, tiền huấn luyện toàn bộ mô hình trên một bộ dữ liệu phong phú ngày càng trở nên phổ biến. Lý tưởng nhất, tiền huấn luyện khiến mô hình phát triển các khả năng và kiến thức có mục đích chung, sau đó có thể được chuyển giao cho các nhiệm vụ tiếp theo. Trong các ứng dụng của học chuyển giao thị giác máy tính [9]–[11], tiền huấn luyện thường được thực hiện thông qua học có giám sát trên một tập dữ liệu được gắn nhãn lớn như ImageNet [12], [13]. Ngược lại, các kỹ thuật hiện đại trong học chuyển tiếp NLP thường được huấn luyện trước bằng cách sử dụng phương pháp học không giám sát trên dữ liệu không được gắn nhãn. Cách tiếp cận này gần đây đã được sử dụng để thu được các kết quả học thuật trong nhiều hệ thống điểm chuẩn NLP phổ biến nhất [14]–[16]. Ngoài sức mạnh thực nghiệm của nó, đào tạo trước không giám sát cho NLP đặc biệt hấp dẫn vì dữ liệu văn bản không gắn nhãn có sẵn trên Internet - ví dụ: Dự án Common Crawl với khoảng 20TB dữ liệu văn bản được trích xuất từ các trang web mỗi tháng. Điều này phù hợp một cách tự nhiên cho mạng nơ-ron, được chứng minh là thể hiện khả năng mở rộng đáng kể, tức là thường có thể đạt được hiệu suất tốt hơn chỉ đơn giản bằng cách đào tạo một mô hình lớn hơn trên tập dữ liệu lớn hơn [17]–[21].

Tài liệu

[1] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1243–1252.

[2] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 655–665. [Online]. Available: https://www.aclweb.org/anthology/P14-1062

[3] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1746–1751. [Online]. Available: https://www.aclweb.org/anthology/D14-1181

[4] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'14. Cambridge, MA, USA: MIT Press, 2014, p. 3104–3112.

[5] P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016, p. 2873–2879.

[6] R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *In Proceedings of EMNLP*, 2013, pp. 1631–1642.

[7] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, pp. 1556–1566.

[8] D. Marcheggiani, J. Bastings, and I. Titov, "Exploiting semantics in neural machine translation with graph convolutional networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 486–492.

[9] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1717–1724.

[10] S. Thrun, L. K. Saul, and B. Schölkopf, *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*. MIT press, 2004, vol. 16.

[11] M. Huh, P. Agrawal, and A. Efros, "What makes imagenet good for transfer learning?" 08 2016.

[12] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, p. 211–252, Dec. 2015. [Online]. Available: https://doi.org/10.1007/s11263-015-0816-y

[14] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Universal Language Model Fine-tuning for Text Classification*, p. 278.

[15] L. Dong, N. Yang, W. Wang, F. Wei, X. Liu, Y. Wang, J. Gao, M. Zhou, and H.-W. Hon, "Unified language model pre-training for natural language understanding and generation," 2019.

[16] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," 2019.

[17] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," 2017.

[18] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," 2017.

[19] R. Jozefowicz, O. Vinyals, M. Schuster, N. Shazeer, and Y. Wu, "Exploring the limits of language modeling," 2016.

[20] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *European Conference on Computer Vision*. Springer, 2018, pp. 185–201.

[21] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.