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

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

Faculty of Information Technology, FPT University Global Education, Ho Chi Minh city, Vietnam

Tóm tắt

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.

Từ khóa: Xử lý ngôn ngữ tự nhiên, Mô hình tiền huấn luyện, Trí tuệ nhân tạo, Học máy.

1. 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 đặc trưng (features) đã học trước đó và tránh bắt đầu lại từ đầu.

*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)

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 xử lý ngôn ngữ tự nhiên, chẳng hạn như mạng nơ-ron tích chập (Convolutional Neural Network) (Gehring et al., 2017; Kalchbrenner et al., 2014; Kim, 2014), mạng nơ-ron hồi quy (Recurrent Neural Network) (Sutskever et al., 2014; Liu et al., 2016), mạng nơ-ron đồ thị (Graph Neural Network) (Socher et al., 2013; Tai et al., 2015; Marcheggiani et al., 2018). Các phương pháp xử lý ngôn ngữ tự nhiên không sử dụng mạng nơ-ron thường chủ yếu dựa vào các tính năng được tạo thủ công rời rạc, trong khi các phương pháp mạng nơ-ron 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 xử lý ngôn ngữ tự nhiên cụ thể. Do đó, các phương pháp thần kinh giúp mọi người dễ dàng phát triển các hệ thống xử lý ngôn ngữ tự nhiên khác nhau.

Gần đây, việc tiền huấn luyện một Mô hình từ bộ dữ liệu đa dạng ngày càng trở nên phổ biến. Một cách lý tưởng, việc tiền huấn luyện (pre-training) này trang bị cho Mô hình các "khả năng" thông dụng cũng như các "kiến thức" để từ đó có thể dùng để chuyển giao cho các tác vụ cụ thể. Trong các ứng dụng của học chuyển giao vào lĩnh vực Thị Giác Máy Tính (Oquab et al., 2014; Thrun et al., 2004; Huh et al., 2016), tiền huấn luyện thường được thực hiện thông qua học có giám sát trên một tập dữ liệu lớn-có gán nhãn- như ImageNet (Deng et al., 2009; Russakovsky et al., 2015). Ở hướng ngược lại, các kỹ thuật hiện đại dùng cho học chuyển giao trong lĩnh vực Xử Lý Ngôn Ngữ Tự Nhiên lại thường thông qua phương pháp học không giám sát trên bộ dữ liệu không-gán-nhãn. Cách tiếp cận này gần đây đã được sử dụng để thu được các kết quả tiên tiến trong hầu hết các chỉ số NLP phổ biến (Kenton and Toutanova, 2018; Dong et al., 2019; Liu et al., 2019). Bên cạnh điểm mạnh về thực nghiệm, tiền huấn luyện không giám sát cho NLP đặc biệt hấp dẫn vì dữ liệu văn bản không gắn nhãn có sẵn rất nhiều 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, một cách tự nhiên lại

rất phù hợp với các mạng nơ-ron, vốn đã thể hiện khả năng mở rộng đáng kể, nghĩa là hiệu suất có thể đẩy lên cao hơn chỉ đơn giản bằng cách đào tạo một mô hình lớn hơn trên tập dữ liệu lớn hơn (Hestness et al., 2017; Shazeer et al., 2017; Józefowicz et al., 2016; Mahajan et al., 2018; Radford et al., 2019)

Tài liệu

Deng, J., Dong, W., Socher, R., Li, L., Kai Li, and Li Fei-Fei (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Dong, L., Yang, N., Wang, W., Wei, F., Liu, X., Wang, Y., Gao, J., Zhou, M., and Hon, H. (2019). Unified language model pre-training for natural language understanding and generation. *CoRR*, abs/1905.03197.

Gehring, J., Auli, M., Grangier, D., Yarats, D., and Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252. PMLR.

Hestness, J., Narang, S., Ardalani, N., Diamos, G. F., Jun, H., Kianinejad, H., Patwary, M. M. A., Yang, Y., and Zhou, Y. (2017). Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409.

Huh, M., Agrawal, P., and Efros, A. A. (2016). What makes imagenet good for transfer learning? *CoRR*, abs/1608.08614.

Józefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *CoRR*, abs/1602.02410.

Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014). A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 655–665, Baltimore, Maryland. Association for Computational Linguistics.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Universal Language Model Fine-tuning for Text Classification*, page 278.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Liu, P., Qiu, X., and Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, page 2873–2879. AAAI Press.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mahajan, D., Girshick, R. B., Ramanathan, V., He, K., Paluri, M., Li, Y., Bharambe, A., and van der Maaten, L. (2018). Exploring the limits of weakly supervised pretraining. In Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y., editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II*, volume 11206 of *Lecture Notes in Computer Science*, pages 185–201. Springer.

Marcheggiani, D., Bastings, J., and Titov, I. (2018). 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)*, pages 486–492.

Oquab, M., Bottou, L., Laptev, I., and Sivic, J. (2014). Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1717–1724.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L.

(2015). Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision*, 115(3):211–252.

Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q. V., Hinton, G. E., and Dean, J. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *In Proceedings of EMNLP*, pages 1631–1642.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Tai, K. S., Socher, R., and Manning, C. D. (2015). 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)*, pages 1556–1566.

Thrun, S., Saul, L. K., and Schölkopf, B. (2004). *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, volume 16. MIT press.