March 2022: From Static Embeddings to Tranformers. Literature List

Younes Samih, David Arps

Version: March 15, 2022

1 Websites, Blog posts, other educational material

- Jay Alammar: The Illustrated BERT. `https://jalammar.github.io/illustrated-bert/` - Blog post that illustrates the functionality of Transformer-based LMs
- Jay Alammar: The Illustrated Transformer. `https://jalammar.github.io/illustrated-transformer/` - Blog post that illustrates transformers in the context of translation. With video!
- Alexander Rush: The Annotated Transformer. `http://nlp.seas.harvard.edu/2018/04/03/attention.html` - Implementation of the original Transformers paper (Vaswani et al., 2017)
- Dive into Deep Learning (Zhang et al., 2020): Great online textbook with code about different Deep Learning topics. `https://d2l.ai`
- Use LMs in your code with the huggingface library: `https://huggingface.co/`

2 Academic papers

2.1 Language Models

- BERT: Devlin et al. (2019)
- RoBERTa: Liu et al. (2019)
- GPT: Radford et al. (2018), GPT-2: Radford et al. (2019), GPT-3: Brown et al. (2020)
- XLNet: Yang et al. (2019)
- DistilBERT: Sanh et al. (2020)

- ELMO: Peters et al. (2018)
- XLM-RoBERTA (multilingual autoencoder): Conneau et al. (2020)

2.2 Benchmarks

- GLUE: Wang et al. (2018)
- SuperGLUE: Wang et al. (2020)

2.3 Interpretability

- Textbook *Explainable Natural Language Processing*: Søgaard (2021)
- Causativity Neurons: Seyffarth et al. (2021)
- NeuroX: Dalvi et al. (2019), `https://github.com/fdalvi/NeuroX`
- Analyzing Individual Neurons in Pre-trained Language Models: Durrani et al. (2020)

3 References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I. and Amodei, D. (2020), Language models are few-shot learners, *in* H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds, 'Advances in Neural Information Processing Systems', Vol. 33, Curran Associates, Inc., pp. 1877–1901.
URL: *https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf*

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L. and Stoyanov, V. (2020), Unsupervised cross-lingual representation learning at scale, *in* 'Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics', Association for Computational Linguistics, Online, pp. 8440–8451.
URL: *https://aclanthology.org/2020.acl-main.747*

Cui, B., Li, Y., Chen, M. and Zhang, Z. (2019), Fine-tune BERT with sparse self-attention mechanism, *in* 'Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)', Association for Computational Linguistics, Hong Kong, China, pp. 3548–3553.
URL: *https://aclanthology.org/D19-1361*

Dalvi, F., Nortonsmith, A., Bau, D. A., Belinkov, Y., Sajjad, H., Durrani, N. and Glass, J. (2019), Neurox: A toolkit for analyzing individual neurons in neural networks, *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* .

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* 'Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)', Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186.
URL: *https://aclanthology.org/N19-1423*

Durrani, N., Sajjad, H., Dalvi, F. and Belinkov, Y. (2020), Analyzing individual neurons in pre-trained language models, *in* 'Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)', Association for Computational Linguistics, Online, pp. 4865–4880.
URL: *https://aclanthology.org/2020.emnlp-main.395*

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.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K. and Zettlemoyer, L. (2018), Deep contextualized word representations, *in* 'Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)', Association for Computational Linguistics, New Orleans, Louisiana, pp. 2227–2237.
URL: *https://aclanthology.org/N18-1202*

Radford, A., Narasimhan, K., Salimans, T. and Sutskever, I. (2018), Improving language understanding by generative pre-training.
URL: *https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035*

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019), Language models are unsupervised multitask learners.
URL: *http://www.persagen.com/files/misc/radford2019language.pdf*

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2020), Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Seyffarth, E., Samih, Y., Kallmeyer, L. and Sajjad, H. (2021), Implicit representations of event properties within contextual language models: Searching for "causativity neurons", *in* 'Proceedings of the 14th International Conference on Computational Semantics (IWCS)', Association for Computational Linguistics, Groningen, The Netherlands (online), pp. 110–120.
URL: *https://aclanthology.org/2021.iwcs-1.11*

Søgaard, A. (2021), Explainable natural language processing, *Synthesis Lectures on Human Language Technologies* 14(3), 1–123.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I. (2017), Attention is all you need, *in* I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf*

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. R. (2020), SuperGLUE: A stickier benchmark for general-purpose language understanding systems, *arXiv preprint 1905.00537* .
URL: *https://arxiv.org/abs/1905.00537*

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O. and Bowman, S. (2018), GLUE: A multi-task benchmark and analysis platform for natural language understanding, *in* 'Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP', Association for Computational Linguistics, Brussels, Belgium, pp. 353–355.
URL: *https://aclanthology.org/W18-5446*

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R. and Le, Q. V. (2019), Xlnet: Generalized autoregressive pretraining for language understanding, *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox and R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 32, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper/2019/file/dc6a7e655d7e5840e66733e9ee67cc69-Paper.pdf*

Zhang, A., Lipton, Z. C., Li, M. and Smola, A. J. (2020), *Dive into Deep Learning*. `https://d2l.ai`.