Exploring segmentation of data for deep learning

Łukasz Dębowski

Institute of Computer Science
Polish Academy of Sciences

ldebowsk@ipipan.waw.pl
https://home.ipipan.waw.pl/l.debowski/

Vector representations of orthographic words and letter ngrams [4] are a fundamental building block of presently used deep learning algorithms such as transformers in natural language processing [5]. The aim of the project is to explore systematically whether the performance of these algorithms can be improved if one considers some more sophisticated methods for segmentation of input data than using orthographic words or letter ngrams. In particular, we propose to look for better data segmentation algorithms in the class of grammar-based compression algorithms. Grammar-based compression algorithms represent the input text as a succinct context-free grammar that generates this text as a unique production [2, 3, 1]. We suppose that this approach could be applied more generally beyond natural language processing—to music, DNA, etc.

References

[1] Charikar, M., Lehman, E., Lehman, A., Liu, D., Panigrahy, R., Prabhakaran, M., Sahai, A., Shelat, A., 2005. The smallest grammar problem. IEEE Transactions on Information Theory 51, 2554–2576.

[2] de Marcken, C. G., 1996. Unsupervised language acquisition. Ph.D. thesis, Massachussetts Institute of Technology.

[3] Kieffer, J. C., Yang, E., 2000. Grammar-based codes: A new class of universal lossless source codes. IEEE Transactions on Information Theory 46, 737–754.

[4] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality. In: 2013 Conference on Neural Information Processing Systems (NIPS).

[5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: 2017 Conference on Neural Information Processing Systems (NIPS).