Badanie segmentacji danych w głębokim uczeniu

Łukasz Dębowski

Institute of Computer Science
Polish Academy of Sciences

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

Wektorowe reprezentacje słów ortograficznych bądź ngramów liter [4] stanowią podstawowy składnik obecnie stosowanych algorytmów głębokiego uczenia takich jak transformery w przetwarzaniu języka naturalnego [5]. Celem projektu jest systematyczne zbadanie, czy akuratność tych algorytmów można poprawić, jeśli zastosuje się bardziej wyrafinowane metody segmentacji danych wejściowych niż użycie słów ortograficznych lub ngramów liter. W szczególności proponowane jest poszukiwanie lepszych algorytmów segmentacji danych w klasie algorytmów kompresji opartych na gramatykach. Algorytmy kompresji oparte na gramatykach przedstawiają tekst wejściowy jako zwięzłą gramatykę bezkontekstową, która generuje ten tekst jako jedyną produkcję [2, 3, 1]. Przypuszczamy, że to podejście można by zastosować bardziej ogólnie poza przetwarzaniem języka naturalnego — do muzyki, DNA itp.

Literatura

[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).