What If We Encoded Words as Matrices and Used Matrix Multiplication as Composition Function?

Presentation of work originally published in the Proc. of the International Conference on Learning Representations 2019

Lukas Galke[1] Florian Mai[2] Ansgar Scherp[3]

Abstract: We summarize our contribution to the International Conference on Learning Representations *CBOW Is Not All You Need: Combining CBOW with the Compositional Matrix Space Model*, 2019. We construct a text encoder that learns matrix representations of words from unlabeled text, while using matrix multiplication as composition function. We show that our text encoder outperforms continuous bag-of-word representations on 9 out of 10 linguistic probing tasks and argue that the learned representations are complementary to the ones of vector-based approaches. Hence, we construct a hybrid model that jointly learns a matrix and a vector for each word. This hybrid model yields higher scores than purely vector-based approaches on 10 out of 16 downstream tasks in a controlled experiment with the same capacity and training data. Across all 16 tasks, the hybrid model achieves an average improvement of 1.2%. These results are insofar promising, as they open up new opportunities to efficiently incorporate order awareness into word embedding models.

Keywords: machine learning; natural language processing; representation learning

Introduction Word embeddings [CW08, Mi13] are celebrated as one of the most impactful contributions from unsupervised representation learning to natural language processing [Go16]. After unsupervised learning from a large textual corpus, the word embeddings can be transferred to various downstream tasks. Sentence representations are then composed of the sum or the mean of the words in the sentence, the so-called continuous bag-of-words [Mi13]. Since these operations are inherently commutative, any information of word order is lost. For instance, the following two sentences would yield the exact same embedding: "The movie was not awful, it was rather great." and "The movie was not great, it was rather awful." A classifier based on the continuous bag-of-words embedding of these sentences would inevitably fail to distinguish the two different meanings [Go17, p. 151]. While using n-grams is a common choice to bring order-awareness into traditional classifiers, storing embeddings for all n-gram combinations would require exponential space. Other approaches such as contextualized word representations [Pe18] require substantially more parameters. We identify the need for efficient, order-aware, word embedding models.

[1] Kiel University / ZBW, Germany lga@informatik.uni-kiel.de
[2] Idiap Research Institute, Martigny, Switzerland florian.mai@idiap.ch
[3] University of Essex, United Kingdom ansgar.scherp@essex.ac.uk

doi:10.18420/inf2019_47

Approach We propose to encode each word as a matrix and to use matrix multiplication as composition function. Because of the associative property, merely $O(\log n)$ sequential steps are sufficient to encode a sentence. Frequent n-grams can be precomputed via dynamic programming. The idea was theoretically explored earlier by Rudolph and Giesbrecht [RG10] as the compositional matrix space model of language without providing any learning algorithm. We show that the CBOW training objective [Mi13] can be adapted to obtain an unsupervised and efficient training scheme by making two adaptions: On the one hand, we modify the initialization scheme such that the expected value of chained matrix multiplications is constant. On the other hand, we chose a random word as target instead of the center word to alleviate bias.

Results and Conclusion Our experiments [MGS19] show that matrix-based embeddings yield an increase in 9 out of 10 linguistic probing tasks compared to vector-based embeddings. We find that matrix-based and vector-based models complement each other well. When training a joint model with both matrix- and a vector-based components, the model yields an increased performance on 10 out of 16 downstream tasks compared the vector-based approach trained on the same data with the same capacity. The average improvement across all 16 tasks is 1.2%. These results are insofar promising, as they open up new opportunities to efficiently incorporate order awareness into word embedding models.

Acknowledgement This research was supported by the Swiss National Science Foundation under grant number "FNS-30216".

References

[CW08] Collobert, Ronan; Weston, Jason: A unified architecture for natural language processing: deep neural networks with multitask learning. In: ICML. ACM, 2008.

[Go16] Goth, Gregory: Deep or shallow, NLP is breaking out. Commun. ACM, 59(3), 2016.

[Go17] Goldberg, Yoav: Neural Network Methods for Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2017.

[MGS19] Mai, Florian; Galke, Lukas; Scherp, Ansgar: CBOW Is Not All You Need: Combining CBOW with the Compositional Matrix Space Model. In: International Conference on Learning Representations. 2019.

[Mi13] Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Gregory S.; Dean, Jeffrey: Distributed Representations of Words and Phrases and their Compositionality. In: NIPS. 2013.

[Pe18] Peters, Matthew E.; Neumann, Mark; Iyyer, Mohit; Gardner, Matt; Clark, Christopher; Lee, Kenton; Zettlemoyer, Luke: Deep Contextualized Word Representations. In: NAACL-HLT. Association for Computational Linguistics, 2018.

[RG10] Rudolph, Sebastian; Giesbrecht, Eugenie: Compositional Matrix-Space Models of Language. In: ACL. The Association for Computer Linguistics, 2010.