The Journal of Open Source Software

Embeddings.jl: easy access to pretrained word embeddings from Julia

Lyndon White[1] **and David Ellison**[2]

1 The University of Western Australia **2** None

DOI: 10.21105/joss.01013

Software

- Review
- Repository
- Archive

Submitted: 31 August 2018
Published: 10 April 2019

Summary

Embeddings.jl is a tool to help users of the Julia programming language (Bezanson, Edelman, Karpinski, & Shah, 2017) make use of pretrained word embeddings for natural language processing. Word embeddings are a very important feature representation in natural language processing. The use of embeddings pretrained on very large corpora can be seen as a form of transfer learning. It allows knowledge of lexical semantics derived from the distributional hypothesis– that words occurring in similar contexts have similar meaning– to be injected into models which may have only limited amounts of supervised, task oriented training data.

Many creators of word embedding methods have generously made sets of pretrained word representations publicly available. Embeddings.jl exposes these as a standard matrix of numbers and a corresponding array of strings. This lets Julia programs use word embeddings easily, either on their own or alongside machine learning packages such as Flux (Innes, 2018). In such deep learning packages, it is common to use word embeddings as an input layer of a LSTM (long short term memory) network or other machine learning model, where they may be kept invariant or used as initialization for fine-tuning on the supervised task. They can be summed to represent a bag of words, concatenated to form a matrix representation of a sentence or document, or used otherwise in a wide variety of natural language processing tasks.

Embeddings.jl makes use of DataDeps.jl (White, Togneri, Liu, & Bennamoun, 2018), to allow for convenient automatic downloading of the data when and if required. It also uses the DataDeps.jl prompt to ensure the user of the embeddings has full knowledge of the original source of the data, and which papers to cite etc.

It currently provides access to:

- multiple sets of word2vec embeddings (Mikolov, Chen, Corrado, & Dean, 2013) for English
- multiple sets of GLoVE embeddings (Pennington, Socher, & Manning, 2014) for English
- multiple sets of FastText embeddings (Bojanowski, Grave, Joulin, & Mikolov, 2017; Grave, Bojanowski, Gupta, Joulin, & Mikolov, 2018) for several hundred languages

It is anticipated that as more pretrained embeddings are made available for more languages and using newer methods, the Embeddings.jl package will be updated to support them.

References

Bezanson, J., Edelman, A., Karpinski, S., & Shah, V. (2017). Julia: A fresh approach to numerical computing. *SIAM Review*, *59*(1), 65–98. doi:10.1137/141000671

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146. doi:10.1162/tacl_a_00051

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the international conference on language resources and evaluation (lrec 2018)*. Retrieved from https://arxiv.org/abs/1802.06893

Innes, M. (2018). Flux: Elegant machine learning with julia. *Journal of Open Source Software*. doi:10.21105/joss.00602

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv:1301.3781*. Retrieved from http://arxiv.org/pdf/1301.3781v3

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp 2014)* (pp. 1532–1543). doi:10.3115/v1/d14-1162

White, L., Togneri, R., Liu, W., & Bennamoun, M. (2018). DataDeps.jl: Repeatable Data Setup for Replicable Data Science. *ArXiv e-prints*. Retrieved from https://arxiv.org/pdf/1808.01091.pdf