Recurrent neural networks for newsgroup documents classification

1 Description

Recurrent neural networks [1] have been one of the most extensively applied methods to text processing [2, 3, 4]. In comparison to bag-of-words approaches they are able to exploit the sequential information encoded in the text.

2 Objectives

The goal of the project is to implement an RNN-based model (simple RNN and LSTM neurons can be used) for classifying different documents in 20 classes. The 20-Newsgroups dataset [5, 6, 7] will be used [1]. It is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups.

The student should: 1) Design the RNN-based architecture in keras or tensorflow. 2) Apply the network to solve the supervised classification problem. 3) Validate the model.

As in other projects, a report should describe the characteristics of the design, implementation, and results. A Jupyter notebook should include calls to the implemented function that illustrate the way it works.

3 Suggestions

- Read the relevant bibliography about the application of RNNs and LSTMs to text modeling.
- Read previous work on the application of DNNs to the 20-Newsgroups dataset.
- Create a tensorflow or keras implementation to solve the problem.

References

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[2] Ozan Irsoy and Claire Cardie. Opinion mining with deep recurrent neural networks. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 720–728, 2014.

[3] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of the 2010 Interspeech Conference*, volume 2, page 3, 2010.

[4] Tomáš Mikolov, Stefan Kombrink, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE, 2011.

[5] Ken Lang. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339, 1995.

[1] It is available from `http://qwone.com/%7Ejason/20Newsgroups/`

[6] A. M. Cardoso-Cachopo. *Improving methods for single-label text categorization*. PhD thesis, 2007.

[7] Swapnil Hingmire, Sandeep Chougule, Girish K Palshikar, and Sutanu Chakraborti. Document classification by topic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 877–880. ACM, 2013.