个人资料

皮果提

+ 加关注 发私信

访问： 1224389次
积分： 19134
等级： BLOG 7
排名： 第454名

原创： 104篇
转载： 13篇
译文： 2篇
评论： 750条

原 word2vec 中的数学原理详解（一）目录和前言

标签： word2vec CBOW Skip-gram Hierarchical Softmax Negative Sampling

2014-07-19 22:42 93593人阅读 评论(153) 收藏 举报

分类： 语言模型（9）

word2vec 是 Google 于 2013 年开源推出的一个用于获取 word vector 的工具包，它简单、高效，因此引起了很多人的关注。由于 word2vec 的作者 Tomas Mikolov 在两篇相关的论文 [3,4] 中并没有谈及太多算法细节，因而在一定程度上增加了这个工具包的神秘感。一些按捺不住的人于是选择了通过解剖源代码的方式来一窥究竟，出于好奇，我也成为了他们中的一员。读完代码后，觉得收获颇多，整理成文，给有需要的朋友参考。

相关链接

目 录

§1 前言

word2vec 是 Google 于 2013 年开源推出的一个用于获取 word vector 的工具包, 它简单、高效, 因此引起了很多人的关注. 由于 word2vec 的作者 Tomas Mikolov 在两篇相关的论文 ([3], [4]) 中并没有谈及太多算法细节, 因而在一定程度上增加了这个工具包的神秘感. 一些按捺不住的人于是选择了通过解剖源代码的方式来一窥究竟.

第一次接触 word2vec 是 2013 年的 10 月份, 当时读了复旦大学郑骁庆老师发表的论文 [7], 其主要工作是将 SENNA 的那套算法 ([8]) 搬到中文场景. 觉得挺有意思, 于是做了一个实现 (可参见 [20]), 但苦于其中字向量的训练时间太长, 便选择使用 word2vec 来提供字向量, 没想到中文分词效果还不错, 立马对 word2vec 刮目相看了一把, 好奇心也随之增长.

后来, 陆陆续续看到了 word2vec 的一些具体应用, 而 Tomas Mikolov 团队本身也将其推广到了句子和文档 ([6]), 因此觉得确实有必要对 word2vec 里的算法原理做个了解, 以便对他们的后续研究进行追踪. 于是, 沉下心来, 仔细读了一回代码, 算是基本搞明白里面的做法了. 第一个感觉就是, “明明是个很简单的浅层结构, 为什么被那么多人沸沸扬扬地说成是 Deep Learning 呢?”

解剖 word2vec 源代码的过程中, 除了算法层面的收获, 其实编程技巧方面的收获也颇多. 既然花了功夫来读代码, 还是把理解到的东西整理成文, 给有需要的朋友提供点参考吧.

在整理本文的过程中, 和深度学习群的群友 北流浪子 ([15, 16]) 进行了多次有益的讨论, 在此表示感谢. 另外, 也参考了其他人的一些资料, 都列在参考文献了, 在此对他们的工作也一并表示感谢.

参考文献

[1] David E Rumelhart, Geoffrey E Hintont, and Ronald J Williams. **Learning representations by backpropagating errors.** Nature, 323(6088):533-536, 1986.

[2] Yoshua Bengio, Rejean Ducharme, Pascal Vincent, and Christian Jauvin. **A neural probabilistic language model.** Journal of Machine Learning Research (JMLR), 3:1137-1155, 2003.

[3] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. **Efficient Estimation of Word Representations in Vector Space.** arXiv:1301.3781, 2013.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. **Distributed Representations of Words and Phrases and their Compositionality.** arXiv:1310.4546, 2013.

[5] Tomas Mikolov, Quoc V. Le, Ilya Sutskever. **Exploiting Similarities among Languages for Machine Translation.** arXiv:1309.4168v1, 2013.

[6] Quoc V. Le, Tomas Mikolov. **Distributed Representations of Sentences and Documents.** arXiv:1405.4053, 2014.

[7] Xiaoqing Zheng, Hanyang Chen, Tianyu Xu. **Deep Learning for Chinese Word Segmentation and POS tagging.** Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 647-657.

[8] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu and Pavel Kuksa. **Natural Language Processing (Almost) from Scratch.** Journal of Machine Learning Research (JMLR), 12:2493-2537, 2011.

[9] Michael U Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. The Journal ofMachine Learning Research, 13:307-361, 2012.

[10] 百度百科中的“哈夫曼树”词条.

[11] 吴军.《**数学之美**》. 人民邮电出版社, 2012.

[12] http://ml.nec-labs.com/senna/

[13] http://www.looooker.com/archives/5621

[14] licstar. **Deep Learning in NLP (一) 词向量和语言模型**.
http://licstar.net/archives/328

[15] **深度学习 word2vec 笔记之基础篇**.
http://blog.csdn.net/mytestmy/article/details/26961315

[16] 深度学习 word2vec 笔记之算法篇.
http://blog.csdn.net/mytestmy/article/details/26969149

[17] 邓澍军, 陆光明, 夏龙. Deep Learning 实战之 word2vec, 2014.

[18] 杨超. Word2Vec 的一些理解.
http://www.zhihu.com/question/21661274/answer/19331979

[19] 基于权值的微博用户采样算法研究.
http://blog.csdn.net/itplus/article/details/9079297

[20] 利用 word2vec 训练的字向量进行中文分词.
http://blog.csdn.net/itplus/article/details/17122431

[21] Yoav Goldberg, Omer Levy. word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. arXiv: 1402.3722v1, 2014. (http://arxiv.org/pdf/1402.3722v1.pdf)

上一篇 一种并行随机梯度下降法
下一篇 word2vec 中的数学原理详解（二）预备知识

相关文章推荐

- word2vec 中的数学原理详解（一）目录和前言
- 【直播】70天软考冲刺计划--任铄
- word2vec原理解析
- 【直播】打通Linux脉络 进程、线程、调度--宋宝华
- word2vec中的数学原理详解
- 【直播】机器学习之凸优化--马博士
- word2vec 中的数学原理详解（二）预备知识
- 【套餐】MATLAB基础+MATLAB数据分析与统计-…
- word2vec数学原理
- 【课程】3小时掌握Docker最佳实战--徐西宁
- word2vec 中的数学原理详解（五）基于 Negative…
- 【课程】深度学习基础与TensorFlow实践--AI100
- Word2Vec数据集
- word2vec 中的数学原理详解（四）基于 Hierarchi…
- word2vec工具下载
- word2vec 中的数学原理详解