文本分类算法的实现与验证

钟函汛
hanxun_zhong@ruc.edu.cn

摘要

本文实现并验证了 RNN, TextRCNN, Transformer 等 3 种文本分类算法。在垃圾短信分类任务上进行了实验。实验结果表明，TextRCNN 在准确率及 F1 指标上超过了其余基准模型。

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

文本分类, 深度学习

ACM Reference Format:

钟函汛. 2018. 文本分类算法的实现与验证. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 介绍

分类 (Classification) 是指自动对数据进行标注。人们在日常生活中通过经验划分类别。但是要依据一些规则手工地对互联网上的每一个页面进行分类，是不可能的。因此，基于计算机的高效自动分类技术成为人们解决互联网应用难题的迫切需求。与分类技术类似的是聚类，聚类不是将数据匹配到预先定义的标签集合，而是通过与其他数据相关的隐含结构自动的聚集为一个或多个类别。文本分类是数据挖掘和机器学习领域的一个重要研究方向。

分类是信息检索领域多年来一直研究的课题，一方面以搜索的应用为目的来提高有效性和某些情况下的效率；另一方面，分类也是经典的机器学习技术。在机器学习领域，分类是在有标注的预定义类别体系下进行，因此属于有监督的学习问题；相反聚类则是一种无监督的学习问题。

文本分类 (Text Classification 或 Text Categorization，TC)，或者称为自动文本分类 (Automatic Text Categorization)，是指计算机将载有信息的一篇文本映射到预先给定的某一类别或某几类别主题的过程。文本分类另外也属于自然语言处理领域。本文中文本 (Text) 和文档 (Document) 不加区分，具有相同的意义。

F. Sebastiani 以如下数学模型描述文本分类任务：文本分类的任务可以理解为获得这样的一个函数 $D * C\{T, F\}$，其中，$D = \{d1, d2, , d|D|\}$ 表示需要进行分类的文档，$C = \{c1, c2, , c|C|\}$ 表示预定义的分类体系下的类别集合，T 值表示对于 (dj, ci) 来说，文档 d_j 属于类 c_i，而 F 值表示对于 (dj, ci) 而言文档 d_j 不属于类 c_i。也就是说，文本分类的目标就是要寻找一个有价值的函数映射，准确的完成 $D * C$ 到 T/F 值的函数映射，这个映射过程本质上讲就是所谓的分类器。

在本文中，我们探讨了 RNN，TextRCNN，Transformer 在文本分类任务上的性能。实验结果表明 TextRCNN 在准确率及 F1 指标上超过了其余基准模型。

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

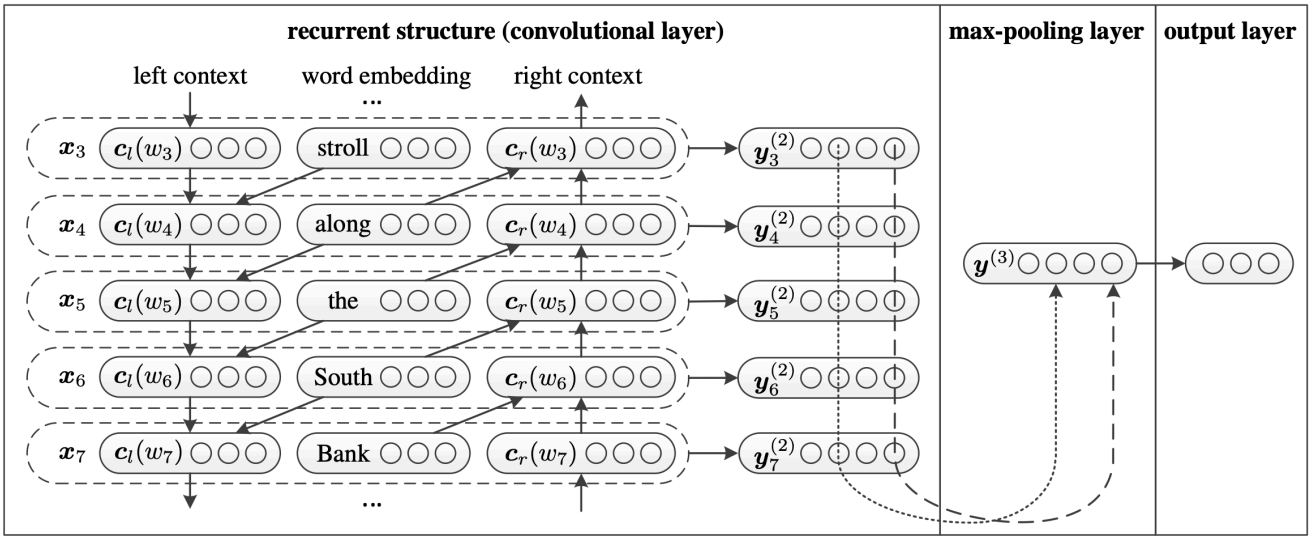

图 1: TextRCNN 结构图

2 相关工作

文本分类作为自然语言处理领域一个基础问题一直受到人们的广泛关注 [1][2][3][5][6][4]。在深度学习方法出现后，由于其结构可以自动获取特征表达能力，从而去掉繁杂的人工特征工程，端到端的解决问题，而得到青睐。[1] 率先利用词向量的平均值通过分类器来预测文本类别。[2] 则利用 CNN 结构捕捉词向量中的重要信息预测结果。[3] 利用 RNN 结构生成获取句向量信息提升了预测效果。由于 RNN 结构会偏向句子中靠后的单词，影响分类效果。[5] 结合了 CNN 与 RNN，将 CNN 中的卷积层替换为 RNN，利用双向循环神经网络捕捉文本信息，再通过 CNN 池化层获得最重要的特征，取得了当时文本分类任务上的最佳成绩。除了利用 CNN 捕获文本中的重要信息，注意力机制也受到了广泛应用。[6] 利用多层注意力结构实现了文本分类。在自注意力网络 [4] 出现后，利用 Transformer 编码器端捕获句子信息用以分类任务取得了非常有竞争性的性能。

3 文本分类模型

3.1 RNN

RNN 模型是序贯的基础模型，我们使用双向 GRU 模型捕捉文本信息，最后选择最后一层隐藏向量拼接后作为文本的特征向量表示，并以此特征向量用以文本预测。

3.2 TextRCNN

RNN 擅长处理序列结构，能够考虑到句子的上下文信息，但 RNN 属于 "biased model"，一个句子中越往后的词重要性越高，这有可能影响最后的分类结果，因为对句子分类影响最大的词可能处在句子任何位置。CNN 属于无偏模型，能够通过最大池化获得最重要的特征，但是 CNN 的滑动窗口大小不容易确定，选的过小容易造成重要信息丢失，选的过大会造成巨大参数空间。

为了解决二者的局限性，[5] 提出了一种新的网络架构，用双向循环结构获取上下文信息，这比传统的基于窗口的神经网络更能减少噪声，而且在学习文本表达时可以大范围的保留词序。其次使用最大池化

层获取文本的重要部分，自动判断哪个特征在文本分类过程中起更重要的作用。

在图 1中展示了 TextRCNN 的总览结构。TextRCNN 将 CNN 中的卷积层替换为 RNN。利用双向循环神经网络捕捉文本信息，再通过 CNN 池化层获得最重要的特征，并利用最终获得的特征向量预测文本类别。

3.3 Transformer

Transformer 模型基于自注意力机制，利用注意力机制替代了 RNN 模型中的循环结构。不仅计算速度更快，而且也更能够捕捉文本信息。基于 Transformer 结构的 Bert 模型在多个自然语言处理任务上均取得了最佳成绩。我们将利用 Transformer 的编码器模块，利用其抽取文本信息，并利用其特征向量预测文本类别。在图 2中，我们展示了 Transformer 模型的具体结构。

4 实验及结果分析

4.1 数据集

我们使用了垃圾短信分类数据集。短信为长度较短的文本。在表 1中我们分别展示了一对文本中的正例与负例。在数据集中，训练集具有 77,584 条数据；验证集含有 10,000 条数据；测试集含有 10,000 条数据。

表 1: 垃圾短信数据集示例

标志	文本
0	其实觉得张家港一个地级市
1	尊敬家长，美琪金榜学校教职工祝全家元宵节快乐，我校开设小学初中高中学科补习班，经典班型對班 x 人班教育专家

4.2 实验结果

在表 2中，我们分别展示了三种模型的实验结果。由实验结果可知，TextRCNN 模型在准确率指标以及 F1 指标上均超过了 RNN 模型与 Transformer 模型，获得了最佳效果。实验证明了利用 CNN 在捕捉文本重要信息上的作用。同时，也证明了在短文本分类任务中，RNN 架构仍然具有较好的实验效果。

表 2: 实验结果

模型	Acc.	F1
RNN	0.9921	0.9958
TextRCNN	**0.9928**	**0.9962**
Transformer	0.9554	0.9770

5 总结与未来工作

我们实现了 RNN、TextRCNN 与 Transformer 三种模型并将其应用在文本分类任务中。在垃圾短信分类数据集上进行了实验，实验结果证明了 TextRCNN 在短文本分类任务上的优越性。在未来工作中，我们将垃圾文本分类器与检索系统结合起来，压低垃圾文本的权重。同时，在垃圾短信分类数据集存在标签不均衡的问题（大部分短信数据为非垃圾短信）。在此后的工作中，将使用数据增强的方法扩充垃圾短信数据来缓解标签不均衡的问题，进一步提高分类器的效果。

REFERENCES

[1] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, Mirella Lapata, Phil Blunsom, and Alexander Koller (Eds.). Association for Computational Linguistics, 427–431. https://doi.org/10.18653/v1/e17-2068

[2] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 1746–1751. https://doi.org/10.3115/v1/d14-1181

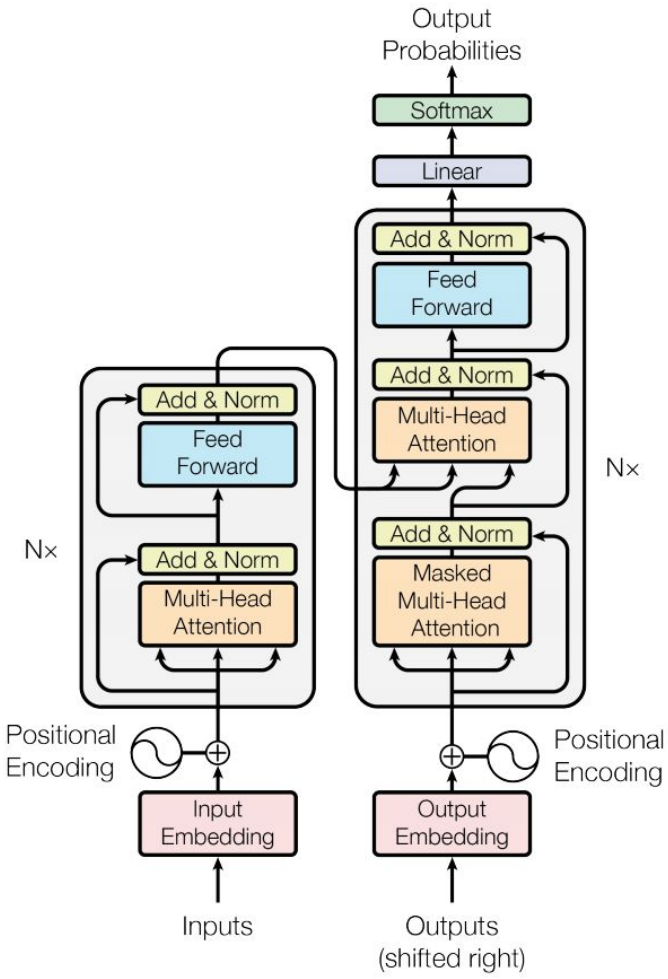

图 2: Transformer 结构图

[3] Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent Neural Network for Text Classification with Multi-Task Learning. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 2873–2879. http://www.ijcai.org/Abstract/16/408

[4] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need

[5] Ruishuang Wang, Zhao Li, Jian Cao, Tong Chen, and Lei Wang. 2019. Convolutional Recurrent Neural Networks for Text Classification. In *International Joint Conference on Neural Networks, IJCNN 2019 Budapest, Hungary, July 14-19, 2019*. IEEE, 1–6. https://doi.org/10.1109/IJCNN.2019.8852406

[6] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alexander J. Smola, and Eduard H. Hovy. 2016. Hierarchical Attention Networks for Document Classification. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, Kevin Knight, Ani Nenkova, and Owen Rambow (Eds.). The Association for Computational Linguistics, 1480–1489. https://doi.org/10.18653/v1/n16-1174