Homework 5

May 15, 2019

[IE801/KSE801] Special Topics in Industrial Engineering II <Machine Learning for Knowledge Service>
TA Sangmook Kim, Jaehoon Oh, Taehyeon Kim, Gihun Lee

HW Description

In this homework, you will analyze sentiment based on movie reviews. You have to implement Recurrent Neural Networks (RNNs) from scratch. We do not give any code, **but you can refer github code or other kernels on Kaggle with reference**. This homework consists of following steps:

1. Download dataset from Kaggle

For this homework, we will use movie reviews dataset from Kaggle. You have to sign up to download dataset, and the link is https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews.

Dataset Description The Rotten Tomatoes movie review dataset is a corpus of movie reviews used for sentiment analysis, originally collected by Pang and Lee [5]. In their work on sentiment treebanks, Socher et al. [8] used Amazon's Mechanical Turk to create fine-grained labels for all parsed phrases in the corpus. This competition presents a chance to benchmark your sentiment-analysis ideas on the Rotten Tomatoes dataset. You are asked to label phrases on a scale of five values: negative, somewhat negative, neutral, somewhat positive, positive. Obstacles like sentence negation, sarcasm, terseness, language ambiguity, and many others make this task very challenging.

2. Data preprocessing

Before implementing networks, you have to go through two steps to preprocess:

Text Preprocessing Text preprocessing is a technique to refine raw text: Tokenization, Capitialization, Removal stopwords, Stemmazation, and so on. As you know, it might have a big impact on your performance. You can find the details in https://towardsdatascience.com/pre-processing-in-natural-language-machine-learning-898a84b8bd47. The python package *nltk* [3] will help your text preprocessing.

Embedding Embedding is to change from text to vector such as **word2vec** [4], **Glove** [6], **FastText** [2]. If you have difficulties in making embedding function yourself, try other Python packages such as *Gensim* [7], and *Spacy* [1] to help this process. You can also implement text preprocessing using these packages.

3. Recurrent Neural Networks (RNNs)

A recurrent neural network is a class of artificial neural network, which is the most proper to train temporal sequence data. Therefore, RNNs have state-of-the-art performance in many Natural Language Process tasks.

For this homework, many-to-one RNN architecture will be proper as following:

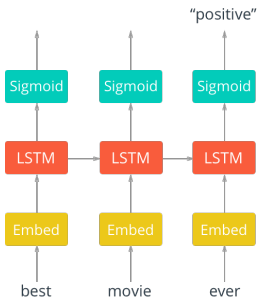

Figure 1: An example of sentiment analysis using RNN

For your better understanding, visit this site: https://towardsdatascience.com/sentiment-analysis-using-rnns-lstm-60871fa6aeba.

Caution: Pytorch is only allowed for this homework.

4. Report

Model description You should describe both a way of text preprocessing and the description of model that you used in details with reasonable reasons. These are what we expect you to describe:

- A way of text preprocessing
- A model description
- scheme changes for hyperparameters
- analysis of results

Score After you train your model, you can check your model performance by submitting on Kaggle. Your submission score will affect your this homework score. If you get score with greater than 0.6, it is okay. Otherwise, there will be a slight penalty based on your score. PLEASE ATTACH YOUR SUBMISSION SCORE including your ID like below:

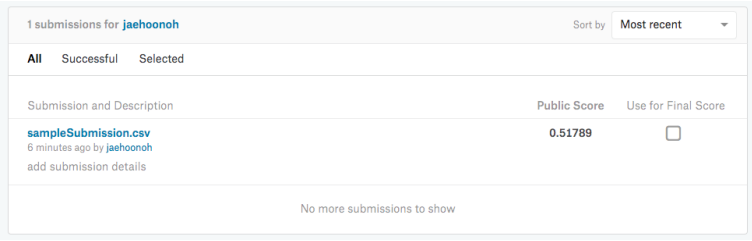

Figure 2: An example of submission capture

Homework 5

May 15, 2019

About the Submission

- The deadline for submission is **23:59 on 29 May (Wed)**, and late submission is not permitted.
- The report should be no more than **3 pages** excluding code.
- You have to submit **.zip file** including both **.ipynb file** and **.pdf file**.
  (Please convert .doc file to **.pdf file**)
- File name should be **[hw5]student_ID.zip** (e.g., [hw5]20191234.zip)
  (If you do not keep this naming, there will be a disadvantage.)
- If you have a question about hw5, **please use Lecture Q&A on KLMS.**

References

[1] Matthew Honnibal and Mark Johnson. An improved non-monotonic transition system for dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1373–1378, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*, 2016.

[3] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[5] Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 115–124. Association for Computational Linguistics, 2005.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *In EMNLP*, 2014.

[7] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

[8] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.