Proposal : Research on Text classification

Yifei Xu (304880196), Yaxuan Zhu (704947072), Ruiqi Gao (104864885)
Department of Statistics
University of California, Los Angeles
fei960922@ucla.edu, yaxuanzhu@ucla.edu, ruiqigao@ucla.edu

1 Introduction

Text classification is a common task in natural language processing (NLP). It predict the most relevant label(s) to a given sentenses or paragraphs. It is widely used in tag recommendation, information retrieval, document categorization, etc.

A good representation of text is key to this problem. There are multiple work on this field including [5] [3] [1] [2]. It is intuitively to introduce attention into text representation. Recently, LEAM [4] introduce attention to linear model classification. We are planning to express this technique to more complex network including LSTM.

Both bag of word and word embedding may be tried in experiment. Also, MLP, CNN, RNN, LSTM and other network structures will be tested.

2 Experiment Expectation

2.1 Dataset

We are planning to use one of these public dataset,

- Reuters Newswire Topic Classification (Reuters-21578). A collection of news documents that appeared on Reuters in 1987 indexed by categories. It has 90 classes, 7769 training documents and 3019 testing documents.
- The Extreme Classification Repository: Multi-label Datasets & Code.
- IMDB Movie Review Sentiment Classification (stanford). A collection of movie reviews from the website imdb.com and their positive or negative sentiment.
- News Group Movie Review Sentiment Classification (cornell). A collection of movie reviews from the website imdb.com and their positive or negative sentiment.

First two are multi-label dataset. The rest are single-label dataset.

2.2 Results

We will compare accuracy for multiple methods. AUC curve may provided.

3 Appendix

3.1 Collaboration

Everything is a plann and not decided yet.

- Yifei Xu : Writing proposal, implement core code.

- Yaxuan Zhu : Writing report, implement add-on, optimizing code.
- Ruiqi Gao : Presentation, Dataset preprocessing, model training.

3.2 Timetable

Everything is a plann and not decided yet.

- Apr.19th : Finish Proposal
- May 15th : Finish core part
- May 30th : Finish alpha version code
- Jun.7th : Finalize code, finish report and presentation

Acknowledgments

This template is NIPS2017 template downloaded from NIPS2017 website.

This proposal is for 2019 Spring CS269 Project. NO DISTRIBUTION.

References

[1] Hareesh Bahuleyan, Lili Mou, Olga Vechtomova, and Pascal Poupart. Variational attention for sequence-to-sequence models. *arXiv preprint arXiv:1712.08207*, 2017.

[2] Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. Latent alignment and variational attention. In *Advances in Neural Information Processing Systems*, pages 9712–9724, 2018.

[3] Yue Jiao, Jonathon Hare, and Adam Prügel-Bennett. Probabilistic semantic embedding, 2019.

[4] Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. Joint embedding of words and labels for text classification. *arXiv preprint arXiv:1805.04174*, 2018.

[5] Ronghui You, Suyang Dai, Zihan Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Extreme multi-label text classification with multi-label attention based recurrent neural networks. *arXiv preprint arXiv:1811.01727*, 2018.