Artificial Neural Network based Self-taught Learning for text classification

0 Abstract

1 Introduction

2 Related work

3 Self-taught learning Model

4 Unsupervised learning

5 Supervised learning

6 Experiment

7 Conclusion

3 Self-taught learning Model

Labeled data is precious for machine learning and pattern recognition especially for data-driven approaches like deep learning. It is often very difficult and expensive to obtain labeled data. In this paper, we present a neural network based self-taught learning model. We apply CNN based classification on limited number of labeled data. Then we enhance this classification through randomly downloaded text from internet.

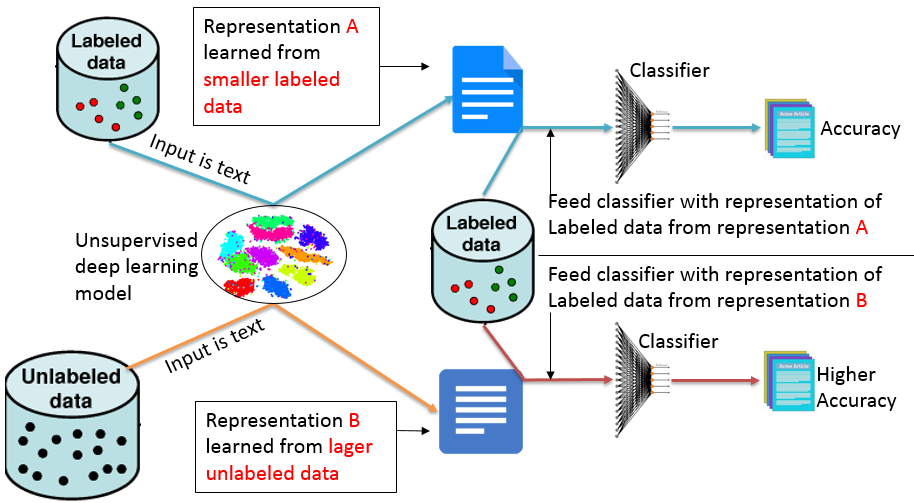


Figure 1: Self-taught learning model diagram

As Figure 1, Labeled data is the text with labels ready for classification. Unlabeled data is the text randomly downloaded from internet.

First, we let the limited labeled data go through our unsupervised deep learning model, then representation A is learned for each word in text. Then we feed the classifier with representation A of labeled data. The result will be Accuracy RA of classification.

Second, let the unlabeled data go through our unsupervised deep learning model, then representation B is learned for each word in text. Then we feed the classifier with representation B of labeled data. The result will be Accuracy RB of classification.

If accuracy RB is higher than RA, we can say the randomly downloaded text really enhanced our limited labeled data.

Unsupervised learning model

Self-taught learning through Autoencoder

Self-taught learning through Word2vec We use datasets S1, S2, S3, and S4 to train the Word2vec model respectively, so we get word representations from the trained model. These representations are ready for our classification tasks on smaller labeled datasets.

Self-taught learning through Doc2vec

We use datasets S1, S2, S3, and S4 to train the Word2vec model respectively, so we get document representations from the trained model. These representations are ready for our classification tasks on smaller labeled datasets.

5 Experiments

Datasets

We utilize four datasets which have been widely used in text classification. The Source datasets which are treated as unlabeled data is shorten as S. and the Target datasets with labels are shorted as T.

Following are all datasets used in our experiment.

|  |  |  |  |
| --- | --- | --- | --- |
| Datasets | Description | Labels | From |
| BBC News | Consists of 2225 documents from the [BBC](http://news.bbc.co.uk/) news website corresponding to stories in five topical areas from 2004-2005 | Class Labels: 5 (business, entertainment, politics, sport, tech) | http://mlg.ucd.ie/datasets/bbc.html |
| Common Crawl news dataset | 27,000 documents for news articles | Class Labels: 5 (business, entertainment, politics, sport, tech) | http://commoncrawl.org/ |
| Cornell Movie Review Data | 1000 positive and 1000 negative processed reviews. Introduced in Pang/Lee ACL 2004. Released June 2004. | Class Labels: 2 (positive and negative) | http://www.cs.cornell.edu/people/pabo/movie-review-data/ |
| Stanford Large Movie Review Dataset | 50,000 highly polar movie reviews | Class Labels: 2 (positive and negative) | http://ai.stanford.edu/~amaas/data/sentiment/ |
| Amazon Fine Food Review dataset | 500,000 food reviews from Amazon | Class Labels: 2 (positive and negative) | https://www.kaggle.com/snap/amazon-fine-food-reviews |

From the above datasets, we use all 2,225 articles of BBC News dataset as target dataset T1 and all 27,000 articles of Common Crawl news as source dataset S1, which is treated as unlabeled articles. Then we use all 2000 reviews of Cornell Movie Review Dataset as target dataset T2 and 25,000 reviews of Stanford large movie review dataset as source dataset S2, which is treated as unlabeled reviews. Then we use 10,000 reviews of Stanford large movie review dataset as target dataset T3 and another 25,000 reviews of Stanford large movie review dataset as source dataset as S3, which is treated as unlabeled reviews. Finally, we use 10,000 reviews of Amazon fine food review dataset as target dataset T4 and another 25,000 reviews of Amazon fine food review dataset as source dataset S4, which is treated as unlabeled data.

Following shows the datasets for each experiment:

|  |  |  |  |
| --- | --- | --- | --- |
| Experiments | Source Datasets | Function | Target Datasets |
| Experiment 1 | S1 | Enhance classification of | T1 |
| Experiment 2 | S2 | Enhance classification of | T2 |
| Experiment 3 | S3 | Enhance classification of | T3 |
| Experiment 4 | S4 | Enhance classification of | T4 |

Experiment process:

Since we try to use downloaded unlabeled data to better understand our existing labeled data. We Word2vector to represent each word in text, then we feed the vectors to CNN model for classification.

5 Conclusion

References:

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." *nature* 521.7553 (2015): 436.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in neural information processing systems*. 2012.

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Chen, Yu, and Mohammed J. Zaki. "Kate: K-competitive autoencoder for text." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2017.