[1] Deep learning, Yann LeCun

0 Abstract

Deep learning has good result and ….. in the paper deep learning[1]. Comparing to traditional machine learning, deep learning is data-driven machine learning instead of model driven. So, data is the most important for deep learning. We can download unlabeled data from all domain, However, the data with high quality is hard to achieve, especially, the data with labels are hard to get. As the Self-taught learning: Transfer Learning from Unlabeled Data [2] described, Self-taught Learning could use easily obtained unlabeled data, enhance model training on small set labeled data. So, we tested on several unsupervised learning methods as approaches to implement self-taught learning on classification task on text data.

1 Introduction

2 Related work

3 Methods

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.

4 Experiments

**Datasets**

We utilize four datasets that have been widely used in text classification. We have Source dataset short as S as well as Target dataset short as T for each group of experiments.

Table 1 provides statistics of all datasets used in our experiment.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Datasets | Description | Labels & number of text document | | Total  articles |
| T1 | BBC news datasets | Business | 502 | 2,098 |
| Entertainment | 364 |
| Politics | 399 |
| Sports | 497 |
| Tech | 336 |
| T2 | Movie Review Datasets from Cornell | Negative | 5,331 | 10,662 |
| Positive | 5,331 |
| T3 | part1 of Movie review datasets From Stanford | Negative | 4,500 | 9,000 |
| Positive | 4,500 |
| T4 | part1 of Amazon food review Datasets | Negative | 5,000 | 10,000 |
| Positive | 5,000 |

Model: our model is CNN text classification. And we are not using google pretrained word2vector, and we use skip-gram trained our own word2vec.

**CNN Model (give more details)**

**Text classification without self-taught learning**

Following is the working flow for text classification without self-taught learning.

Feed target dataset T1 to skip gram model and we get word representation R1 for every word in the target dataset T1.

Feed target dataset T2 to skip gram model and we get word representation R2 for every word in the target dataset T2.

Feed target dataset T3 to skip gram model and we get word representation R3 for every word in the target dataset T3.

Feed target dataset T4 to skip gram model and we get word representation R4 for every word in the target dataset T4.

Then the Target datasets T1, T2, T3 and T4 was randomly split as 9/10 for training set, and 1/10 as test set. With the values in Representation R1, R2, R3, and R4. We feed x of training set to CNN model, with 200 epochs training, we get our model for test. After test data going through the classification model, then we get the accuracy of 90%, 91%, 93%, 96% on test set of T1, T2, T3 and T4 correspondingly.

**Text classification using self-taught learning**

Since labeled data is scare, and unlabeled data can be obtained easily.

5 Conclusion