

Text Classification and Multitask Learning

Hong Zhang

Yahoo Research

October 29, 2017

- 1 Text Classification
- 2 Regularization Techniques
- 3 Discussion

- Text classification is a classical topics in Natural Language Processes, e.g., spam filtering, sentiment analysis, DA, RA, intent models, . . .
- Two components: Sentence representation and classification

Sentence embedding

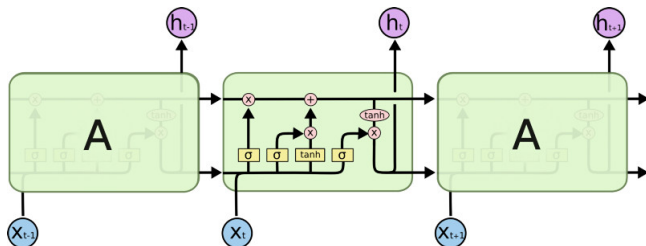
Traditional Methods:

- Bag of words and its TFIDF (term frequency-inverse document frequency)
- Bag of ngrams and its TFIDF

Deep Learning:

- Word-based ConvNets
- Character-based ConvNets
- *LSTM*

LSTM



- Use the last state as the sentence representation
- Use the average of the states as the representation (not yet implemented)

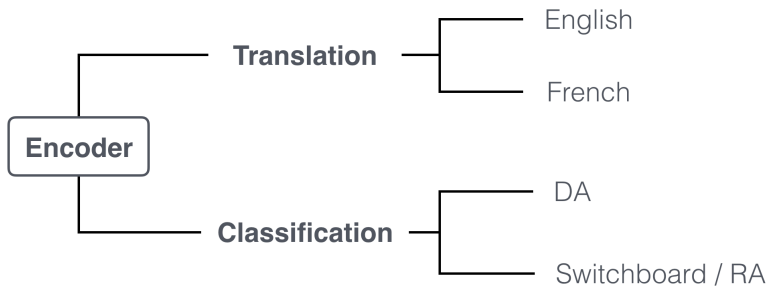
- 1 Text Classification
- 2 Regularization Techniques
- 3 Discussion

Regularizations

DA-movie has 8.5k training data and 1k validation data. The dimension of Glove embedding is 300 and hidden state's dimension is 256. Overfitting is an obvious problem. Standard regularization techniques:

- early stop
- label smoothing
- ridge/lasso
- multitask learning

multitask learning



Discussion

Possible future directions:

- ensemble
- try different classifier, for example SVM
- use the mean of the hidden state instead of the last state
- add more features

Discussion: how much data do we need?

It is difficult to predict the learning curve

- VC dimension provides a theoretical bound, which is useless in this case.
- Depends on a lot of factors:
 - classification method
 - complexity of the classifier
 - how well the classes are separated
 - data quality
 - ...

Dataset used by LeCun et al. (Character-based ConvNet)

| Dataset | Classes | Train Samples | Test Samples | Epoch Size |
|------------------------|---------|---------------|--------------|------------|
| AG's News | 4 | 120,000 | 7,600 | 5,000 |
| Sogou News | 5 | 450,000 | 60,000 | 5,000 |
| DBPedia | 14 | 560,000 | 70,000 | 5,000 |
| Yelp Review Polarity | 2 | 560,000 | 38,000 | 5,000 |
| Yelp Review Full | 5 | 650,000 | 50,000 | 5,000 |
| Yahoo! Answers | 10 | 1,400,000 | 60,000 | 10,000 |
| Amazon Review Full | 5 | 3,000,000 | 650,000 | 30,000 |
| Amazon Review Polarity | 2 | 3,600,000 | 400,000 | 30,000 |

Error Rate

| Model | AG | Sogou | DBP. | Yelp P. | Yelp F. | Yah. A. | Amz. F. | Amz. P. |
|--------------------|--------------|--------------|-------------|--------------|--------------|--------------|--------------|--------------|
| BoW | 11.19 | 7.15 | 3.39 | 7.76 | 42.01 | 31.11 | 45.36 | 9.60 |
| BoW TFIDF | 10.36 | 6.55 | 2.63 | 6.34 | 40.14 | 28.96 | 44.74 | 9.00 |
| ngrams | 7.96 | 2.92 | 1.37 | 4.36 | 43.74 | 31.53 | 45.73 | 7.98 |
| ngrams TFIDF | 7.64 | 2.81 | 1.31 | 4.56 | 45.20 | 31.49 | 47.56 | 8.46 |
| Bag-of-means | 16.91 | 10.79 | 9.55 | 12.67 | 47.46 | 39.45 | 55.87 | 18.39 |
| LSTM | 13.94 | 4.82 | 1.45 | 5.26 | 41.83 | 29.16 | 40.57 | 6.10 |
| Lg. w2v Conv. | 9.92 | 4.39 | 1.42 | 4.60 | 40.16 | 31.97 | 44.40 | 5.88 |
| Sm. w2v Conv. | 11.35 | 4.54 | 1.71 | 5.56 | 42.13 | 31.50 | 42.59 | 6.00 |
| Lg. w2v Conv. Th. | 9.91 | - | 1.37 | 4.63 | 39.58 | 31.23 | 43.75 | 5.80 |
| Sm. w2v Conv. Th. | 10.88 | - | 1.53 | 5.36 | 41.09 | 29.86 | 42.50 | 5.63 |
| Lg. Lk. Conv. | 8.55 | 4.95 | 1.72 | 4.89 | 40.52 | 29.06 | 45.95 | 5.84 |
| Sm. Lk. Conv. | 10.87 | 4.93 | 1.85 | 5.54 | 41.41 | 30.02 | 43.66 | 5.85 |
| Lg. Lk. Conv. Th. | 8.93 | - | 1.58 | 5.03 | 40.52 | 28.84 | 42.39 | 5.52 |
| Sm. Lk. Conv. Th. | 9.12 | - | 1.77 | 5.37 | 41.17 | 28.92 | 43.19 | 5.51 |
| Lg. Full Conv. | 9.85 | 8.80 | 1.66 | 5.25 | 38.40 | 29.90 | 40.89 | 5.78 |
| Sm. Full Conv. | 11.59 | 8.95 | 1.89 | 5.67 | 38.82 | 30.01 | 40.88 | 5.78 |
| Lg. Full Conv. Th. | 9.51 | - | 1.55 | 4.88 | 38.04 | 29.58 | 40.54 | 5.51 |
| Sm. Full Conv. Th. | 10.89 | - | 1.69 | 5.42 | 37.95 | 29.90 | 40.53 | 5.66 |
| Lg. Conv. | 12.82 | 4.88 | 1.73 | 5.89 | 39.62 | 29.55 | 41.31 | 5.51 |
| Sm. Conv. | 15.65 | 8.65 | 1.98 | 6.53 | 40.84 | 29.84 | 40.53 | 5.50 |
| Lg. Conv. Th. | 13.39 | - | 1.60 | 5.82 | 39.30 | 28.80 | 40.45 | 4.93 |
| Sm. Conv. Th. | 14.80 | - | 1.85 | 6.49 | 40.16 | 29.84 | 40.43 | 5.67 |

Thank you.