Authorship Attribution for Small Documents

BY ROBERT SMITH, MILES BAER, WILLIAM COPE, NATHANIEL BOYER, CHARLES JOHNSON

Text Processing for Small Documents

- Neural Networks are generally good at Classification Problems
- There is a rise of text that is Limited in size
  - Twitter
  - Email
  - Facebook
  - Online Reviews
- Inspiration for Project
  - Convolutional Neural Networks for Authorship Attribution of Short Texts (Shrestha et. Al)

Encoding Methods & Preprocessing

ENCODINGS

- Word2Vec
  - Efficient Estimation of Word Representations in Vector Space
  - Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean
- Doc2Vec
  - Distributed Representations of Sentences and Documents
  - Le, Quoc and Mikolov, T.
- Our Custom Method
  - Simple Sum of Word2Vec representations generated with spaCy and genism

PREPROCESSING

- Reversed the Input Text
- Non-reversed Input Text
- Removing Punctuation

Neural Networks used in the Project

- CNN
  - L1,L2 Regularizers
  - MaxPooling1D
  - Flatten
- Dense
  - Flatten

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_1 (Conv1D)            (None, 298, 300)          1200
_________________________________________________________________
max_pooling1d_1 (MaxPooling1 (None, 149, 300)          0
_________________________________________________________________
flatten_1 (Flatten)          (None, 44700)             0
_________________________________________________________________
dense_1 (Dense)              (None, 300)               13410300
_________________________________________________________________
dense_2 (Dense)              (None, 3)                 903
=================================================================
Total params: 13,412,403
Trainable params: 13,412,403
Non-trainable params: 0
_________________________________________________________________
```

Dataset and Sampling

- 53,000+ Reviews for Office Supplies
- Author Training Size
  - 3
  - 35
  - 500
  - 5000
- Normal Sampling vs Oversampling
  - Using 3 times the training data

Results

Custom Encoding

	3 No Sample	35 No Sample	3 Over Sample	35 Over Sample
CNN_DENSE	12.5%	1.2%	18.7%	0.83%
CNN_DENSE_R	68.7%	2.5%	12.5%	1.6%
SMALL_DENSE	68.7%	0.83%	68.7%	4.1%

Regular Doc2Vec Encoding

	3 No Sample	35 No Sample	3 Over Sample	35 Over Sample
CNN_DENSE	62.5%	7.5%	75.0%	8.3%
CNN_DENSE_R	43.7%	5.0%	75.0%	7.5%
SMALL_DENSE	62.5%	13.3%	43.7%	9.2%

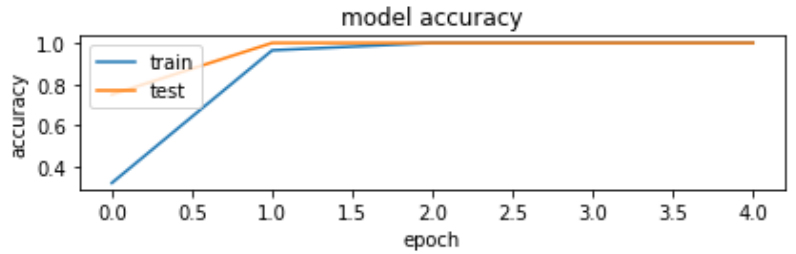
Regular Doc2Vec Encoding

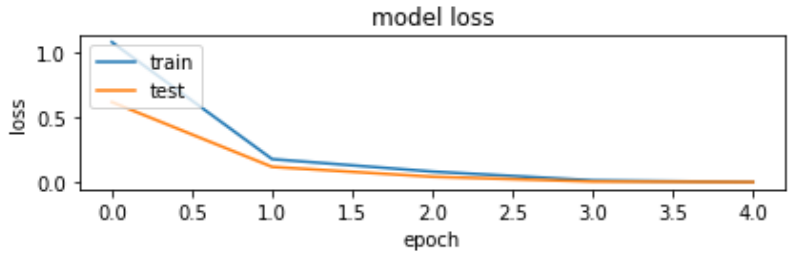
	3 No Sample	35 No Sample	3 Over Sample	35 Over Sample
CNN_DENSE	68.7%	6.2%	50.0%	7.1%
CNN_DENSE_R	43.7%	1.6%	62.5%	1.6%
SMALL_DENSE	56.2%	8.3%	43.7%	43.7%

Results Cont…

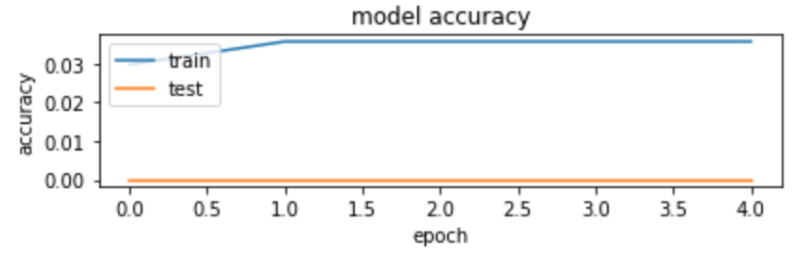
- Regular Doc2Vec Encoding
  - Using duplicate data helped network learn authors better
- Reversed Doc2Vec Encoding
  - Each sentence was reversed before encoding
  - Poor performance when oversampling was used
- Custom Encoding
  - Best results when only 3 authors and normal sample size

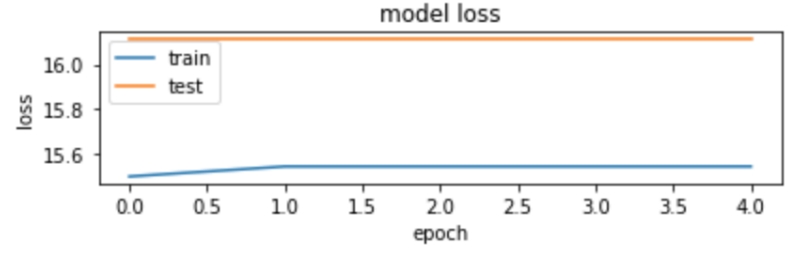
```
16/16 [==============================] - 0s 14ms/step
Test loss: 0.9624202847480774
Test accuracy: 0.75
```






```
239/239 [==============================] - 2s 9ms/step
Test loss: 15.983215874707849
Test accuracy: 0.008368200836820083
```





Conclusion

- Mostly Unsuccessful
  - Doc2Vec
    - Performed the Best (3 Authors, normal encoding)
  - Custom Encoding
    - Performed the Worst
- Overfitting was Prevalent
  - Small Number of Epochs Needed

Future Works

- Explore the Method used in the previously Mention Paper
  - Embedding Layer
  - Concatenated CNNs (filters = [3,4,5])
    - Filters are window/kernel sizes (it will form a numerical representation over n number of words, characters, etc.)
  - Softmax Layer
- Change Activation Methods
- Change Loss Metric
- Adjust Learning Rate
- Mess Around with Punctuation
  - Use punctuation as individual "words"
- More Even Data Distribution
- More Reviews per Author

Questions?

References

- Le, Q. and Mikolov, T., 2014, January. Distributed representations of sentences and documents. In *International Conference on Machine Learning* (pp. 1188-1196).
- *Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR, 2013.*
- Shrestha, P., Sierra, S., Gonzalez, F., Montes, M., Rosso, P. and Solorio, T., 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers* (Vol. 2, pp. 669-674).